INFORMING PUBLIC HEALTH APPROACHES TO OBESITY AND SMOKING USING GENOME-WIDE ASSOCIATION STUDIES: GENETIC EPIDEMIOLOGY AFFIRMS THE IMPORTANCE OF EARLY PREVENTION

Daniel W Belsky

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Health Policy & Management, Gillings School of Public Health.

Chapel Hill
2012

Approved by:

Joseph P Morrissey

Andrea K Biddle

John Vernon

James P Evans

Avshalom Caspi

ABSTRACT

DANIEL W. BELSKY: Informing Public Health Approaches to Obesity and Smoking
Using Genome-Wide Association Studies:
Genetic Epidemiology Affirms the Importance of Early Prevention
(Under the Direction of Joseph P. Morrissey and Avshalom Caspi)

Rapid advances in technology and scientific methods stimulated by the sequencing of the human genome have yielded discoveries that begin to uncover the genetic roots of common chronic health conditions. However, the implications of these discoveries for public health research and practice remain unclear. Three questions are central to building a translational pipeline that links genetic discovery research with interventions to improve health: First, when in the life course do genetic risks become manifest? Second, what are the magnitudes of risks that can be predicted using genetic information? And third, do genetic markers provide new information about risk over and above the existing technology of family health history assessment? This dissertation research seeks to address these questions for two prevalent and costly sources of morbidity and early mortality, obesity and smoking. Results reveal that (1) genetic risks manifest early in the development of obesity and smoking through processes that may be amenable to public health intervention; (2) the magnitudes of risk that can be predicted using genetic information are small; but (3) the risk information provided by genetic markers is independent of information available in a family history. These findings affirm recommendations of caution in the application of genetic information to predict health risks in individuals, but suggest promise as more powerful but less common genetic risks are discovered in the continuing evolution of genomic research. Further, these findings recommend an increased

focus on childhood and adolescence in genetic discovery research and add a genetic rationale to arguments for early intervention to prevent obesity and smoking.

ACKNOWLEDGEMENTS

Thanks to my committee for their hard work and support in mentoring this dissertation. Thanks also to Terrie Moffitt, who provided mentoring and support, and without whom the research presented in this dissertation could not have been conducted. Thanks to Karen Sugden and Benjamin Williams for being my teachers and guides as I learned my way around the world of genetics. Thanks to Sarah Hamilton, David Collins, and Ute Whittmann Pair for grants and administrative support, to Renate Houts for expert statistical consultation and to HonaLee Harrington for superb guidance and support in utilizing the vast data archives of the Dunedin Multidisciplinary Health and Development Study.

In addition to the support of the Moffitt-Caspi Lab, this dissertation was supported by two predoctoral fellowships:

- NICHD Ruth L. Kirschstein National Research Service Award (NRSA) Predoctoral Training Grant (T-32) at the Carolina Consortium for Human Development (CCHD), Eunice Kennedy Shriver National Institute of Child Health and Human Development.
- AHRQ 1R36HS020524-01, F-36 NRSA Predoctoral Fellowship; Agency for Health Care Research & Quality.

I thank the Dunedin Study members, their families, Unit research staff, and Study founder Phil Silva. The Dunedin Multidisciplinary Health and Development Research Unit is supported by the New Zealand Health Research Council. This research received support from UK Medical Research Council grants G0100527 and G0601483, US-National Institute on Aging grant AG032282, and US-NIMH grant MH077874. Additional support was provided by the Jacobs Foundation. The study protocol was approved by the institutional ethical review boards of the participating universities. Study members gave informed consent before participating.

The Atherosclerosis Risk in Communities (ARIC) Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C), R01HL087641, R01HL59367 and R01HL086694; National Human Genome Research Institute contract U01HG004402; and National Institutes of Health contract HHSN268200625226C. The authors thank the staff and participants of the ARIC study for their important contributions. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NIH Roadmap for Medical Research. ARIC data (phs000090.v1.p1) were obtained from dbGaP.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

BMI	Body Mass Index
GRS	Genetic Risk Score
GWAS	Genome-wide Association Study
SNP	Single-nucleotide Polymorphism

CHAPTER 1.

INTRODUCTION

Rapid advances in technology and scientific methods driven by the human genome project have yielded discoveries about the genetic roots of common chronic health conditions.[1] However, the implications of these discoveries for public health research and practice remain unclear.[2] The major engine for genetic discovery in the genomic era has been the genome-wide association study (GWAS). GWAS measure millions of common variants, called single-nucleotide polymorphisms[a] (SNPs), to capture the full range of common variation in the genome. The GWAS "experiment" uses large samples of individuals, often assembled in a case-control design, to test associations between each of these millions of SNPs and a trait or health outcome. The GWAS asks, for each SNP, is the distribution of alleles different in affected cases as compared to unaffected controls? This theory-free data mining approach to genetic discovery allows GWAS to leapfrog current biology.[3] However, the large number of tests (one for each SNP) requires a stringent statistical correction, with the result that GWAS require extremely large samples to discover all but the most powerful genetic risk factors.[4] These large samples are typically assembled, at least in part, from clinical populations. Therefore, follow-up of GWAS discoveries in epidemiologically-sound cohorts is needed to characterize the public health implications of genetic risks.[5] Such characterization is a critical link in the translational pipeline between genetic discovery research and interventions to improve public health.[6] Three questions are central to characterizing GWAS-discovered genetic risks: First, when in the life course do genetic risks become manifest? Second, what are the magnitudes of risks that can be predicted using genetic

[a] The genome is composed of nucleotide chains. Each nucleotide is represented with a letter in the human genetic code. A single nucleotide polymorphism is a single-letter substitution in this code, e.g. from adenine ('A') to guanine ('G') or from cytosine ('C') to tyrosine ('T'), that occurs in at least 1% of the population.

information? and Third, do genetic markers provide new information about risk over and above the existing technology of family health history assessment?

The program of research described in the three empirical studies that comprise this dissertation seeks to characterize the public health implications of GWAS-discovered genetic risks for two prevalent and costly sources of morbidity and early mortality, obesity and smoking. I use theory-free GWAS to derive multi-locus profiles of genetic risk for obesity and smoking, called "genetic risk scores" (GRSs). I validate these GRSs using data from a large, population-based cohort of older adults (n=15,792). I then leverage the power of a complete birth cohort (n=1,037) followed through their fourth decade of life to investigate how genetic risk indexed in the GRSs influences the development of obesity and smoking problems across the first half of the life course. Results reveal that (1) GRSs can be used to investigate GWAS-discovered genetic risks for obesity and smoking in population-based cohorts much smaller than the original GWAS discovery samples; (2) genetic risks identified in GWAS manifest early in the development of obesity and smoking through processes that may be amenable to public health intervention; (3) the magnitudes of risk that can be predicted using genetic information are small; but (4) the risk information provided by genetic markers is independent of information available in a family history. These findings affirm recommendations of caution in the application of genetic information to predict health risks in individuals,[7-9] but suggest promise as more powerful but less common genetic risks are discovered in the continuing evolution of genomic research. Further, findings recommend an increased focus on childhood and adolescence in genetic discovery research and add a genetic rationale to arguments for early intervention to prevent obesity and smoking.[10, 11]

The remainder of this introduction presents (I) The logic for investing GWAS discovered genetic risks in public health research; (II) The methods to be used in this investigation; and (III) The three empirical chapters that follow.

I. Logic for Investigating GWAS-Discovered Genetic Risks in Public health Research

(A) Genetic variation is an important determinant of individual differences in morbidity and mortality. Studies of twins and families indicate that relatives who share more of their genetic code also share liability to many common chronic health conditions.[12] Family studies estimate that as much as 80% of population variation in body mass index and 50% of population variation

in smoking behavior may be attributed to genetics.[13, 14] Recent genetic discoveries appear to explain small fractions of population variance in obesity and smoking.[15] However, these discoveries provide a critical window into how genetic risk operates and, through interactions with other genetic or environmental factors, may account for a larger share of variation in morbidity than initial estimates suggest.[16, 17] Therefore, this research seeks to understand how discovered genetic risks influence the development of obesity and smoking in the population to inform the development of hypotheses about how genetic risk factors with apparently small effects may give rise to large differences in health outcomes.

(B) Genetic information can reveal health risks in time to prevent disease. DNA sequence variants remain constant across the life course; sequence variants that predict health risk in adulthood can be measured accurately from birth to assess risk in pre-symptomatic individuals.[18] Many complex conditions are sensitive to risk exposures early in life and to patterns of health behavior that develop during childhood. This is true of obesity and smoking;[19, 20] and, although evidence-based preventative interventions are available to mitigate early-life risks, poor uptake and adherence pose enduring challenges.[21-25] In addition to improving identification of persons at risk, there is evidence that genetic risk information, if communicated effectively in clinical settings, may help to motivate behavior change.[26, 27] This research attempts to lay a foundation that could ultimately target public health interventions to the most vulnerable "windows" in the development of these health problems.

Genotyping costs are declining rapidly (more than 100,000 fold in the past decade)[28] and personal genomes are making their way to clinicians and, through direct to consumer services, to individuals with no formal medical training.[29] Research is therefore needed to understand the effectiveness of genome-based risk assessments for complex health conditions.[30-32] Ultimately, the effectiveness of genetic risk assessments must be evaluated in terms of their clinical utility—can the results of genetic screens change provider behavior in ways that improve patient health?[33] Before clinical utility can be tested, it is necessary to first generate hypotheses about how genetic information can be effectively deployed.[2] To generate such hypotheses, information is needed about when in the development of a health condition genetic risk becomes manifest, what the magnitudes of that risk are, and whether risk information furnished by genetic markers can also be obtained using the existing technology of family history assessment.[32] The research undertaken in this dissertation seeks to inform the development of

hypotheses about how genetic information can be deployed to prevent obesity and nicotine dependence.

II. Methods for Investigating GWAS-Discovered Genetic Risks in Public health Research

(A) The first step in translating discoveries from the frontiers of genome science into tools that can improve public health is to follow-up discoveries in longitudinal population-based cohort studies. GWAS scan the entire genome for correlations between measures of health and genetic variants, usually single base-pair changes in the human DNA sequence that occur in >1% of the population, called "single nucleotide polymorphisms" (SNPs). GWAS analyses comprise large numbers of statistical tests, one for each SNP measured—a few hundred thousand in early versions, upwards of 2 million in the most recent studies.[4] To minimize the risk of detecting false positive signals, GWAS apply a stringent statistical correction.[b] The field standard significance threshold is p<1x10E-8. At this threshold, only large effects can be detected in standard epidemiological samples.[34] One of the first discoveries of the genomic age was that few SNPs have such large effects.[35] To address the challenge posed in detecting small effects while maintaining the 5% type-1 error rate of conventional epidemiological studies, researchers assembled ever-larger samples.[36] In the case of most complex health conditions, samples of tens of thousands of individuals are needed.[37] These large samples can be effective for discovery research, but follow-up in epidemiologically sound population-based cohorts followed over time is needed to characterize the implications of genetic discoveries for population health.[38]

Population-representative samples are necessary to address selection issues inherent in the case-control designs of many GWAS samples and the recruitment of subjects from clinical populations.[39] Without resolving such issues, it is unclear how genetic risk effects estimated in GWAS translate to the general population.[40] This is a particular problem for obesity and smoking, which are subject to complex environmental influences[41, 42] and are the target of much clinical and public health attention.[43] This makes the clinical populations that constitute significant portions of GWAS samples problematic for interpreting genetic effects. Therefore, this dissertation uses the population-based Atherosclerosis Risk in the Communities (ARIC) cohort[44] to validate genetic risk measurements.

[b] The statistical correction applied in GWAS is a modified version of the Bonferroni correction, which takes the alpha level (acceptable type-1 error rate) and exponentiates it to the power of the number of tests conducted. In GWAS, the Bonferroni correction is adjusted to account for correlation or "linkage" between SNPs.

Longitudinal data including repeated measurements of health states are necessary to address questions of when in the development of a health problem genetic risks become manifest. Questions of when genetic risks becomes manifest are of critical importance for public health practice as intervention to disrupt genetic risk is likely to be most effective before the onset or early in the development of processes that entrain later health problems. Early stages in the development of adult obesity and smoking problems are critical to pathogenesis: Individual differences in obesity risk emerge during gestation and are further established during infancy and childhood through accelerated growth trajectories.[45, 46] Etiological research on smoking highlights the progression from initiation to heavy use during adolescence as a key marker of risk for subsequent dependence.[47, 48] Therefore, this dissertation uses the Dunedin Multidisciplinary Health and Development Study (Dunedin Study) cohort, a complete birth cohort followed over 4 decades with nearly complete retention, to investigate how GWAS-discovered genetic risks for obesity and smoking manifest during development.

(B) Aggregating risk variants discovered in GWAS to create "genetic risk scores" can facilitate investigations of genetic risk in databases with rich longitudinal data to describe the developmental course of a health condition. Prospective longitudinal cohort studies containing repeated measures of health states at multiple points in the life course are necessary to elucidate developmental processes leading to complex diseases.[39] A challenge for research following-up GWAS discoveries in such population-based cohorts is that effect sizes for the individual alleles that are the units of analysis in GWAS are small. In the case of the two conditions examined in this dissertation, obesity and smoking, the most highly penetrant (i.e. most pathogenic) alleles predict at most a half a point increase in adult body-mass index (BMI) or a single cigarette per day increase in tobacco consumption among smokers.[49, 50] A related challenge is that GWAS-identified SNPs do not measure genetic risks with precision.[51] GWAS measure only 100,000 – 2 million of the ~15 million SNPs in the human genome. GWAS methods assume that these SNPs capture the full range of common genomic variation because SNPs that are close together on the genome are non-independent, a phenomenon known as linkage disequilibrium (LD).[52] Two SNPs are said to be in LD when they are inherited together (when their co-occurrence in a population departs significantly from the expectation given their individual frequencies). When LD is strong, SNPs co-occur with sufficient frequency that each SNP serves as a proxy for the other SNP.[53] Proxy SNPs are adequate to test associations in samples of tens of thousands. However, in more modest samples, the imprecision or "noise"

that results from imperfect correlation between a proxy SNP and the causal variant that contributes directly to disease etiology can overwhelm the "signal" and result in failure to detect the association.

Many longitudinal studies with data necessary to investigate the development of complex health conditions are underpowered to test the effects of individual SNPs identified in GWAS.[54] However, there is evidence that many GWAS-identified SNPs contribute additively to disease risk,[55-58] and this is particularly true in the case of obesity and smoking.[59-62] If genetic contributions to risk are additive, it is possible to sum risk alleles across GWAS-identified SNPs to compute a "genetic risk score" (GRS).[63, 64] The resulting GRS provides a valid index of the continuum of genetic risk in the population.[65] Because GRSs measure the aggregate effect of a number of SNPs, they can be used to test associations in the smaller samples that have adequate data to investigate genetic influence over developmental processes. Therefore, as an initial step, this dissertation research uses GWAS discoveries to derive GRSs for obesity and smoking and then uses the GRSs to investigate genetic risk for obesity and smoking.

III. Three Empirical Chapters

(A) The first empirical chapter, *Development and Evaluation of a Genetic Risk Score for Obesity*, describes how results from 16 GWAS of obesity and related phenotypes were used to derive the obesity GRS. The chapter then presents an analysis of the predictive validity of the GRS for adult obesity among European- and African-descent populations in the ARIC cohort.

(B) The second empirical chapter, *Polygenic Risk for Adult Obesity is Mediated by Rapid Childhood Growth*, uses the GRS derived in the first chapter to investigate when genetic risk for obesity manifests in development. This chapter also describes analyses to test whether the obesity GRS provides different information about risk from a family history assessment. This chapter is currently in-press at the *Archives of Pediatrics & Adolescent Medicine*.

(C) The third empirical chapter, *Polygenic Risk Accelerates the Developmental Progression of Smoking Behavior from Initiation to Heavy, Persistent Use and Nicotine Dependence*, uses 3 meta-analyses of GWAS of smoking quantity (cigarettes smoked per-day by individuals who have ever smoked) to derive a GRS for smoking. It then uses this GRS to investigate how genetic risk for smoking relates to developmental and clinical phenotypes of smoking behavior. This

chapter also describes analyses to test whether the smoking GRS provides additional information about risk with that derived simply from a family history assessment.

CHAPTER 2.

DEVELOPMENT AND EVALUATION OF A GENETIC RISK SCORE FOR OBESITY

INTRODUCTION

Genome-wide associations study (GWAS) results represent a potentially rich source of information for etiological and treatment research that builds bridges between genome science and clinical and public health practice. [66, 67] Given the large number of such studies, sufficient GWAS data exist to support such translational research for a number of common chronic health conditions, including obesity [3, 68]. Infrastructure is in place at the start of the translational pipeline with GWAS data banked and curated in continuously updated searchable databases. [3, 69] Likewise, at the other end of the pipeline, evidence from translational research is evaluated to establish the clinical utility of genomic information and to issue guidelines for clinical practice. [6] However, significant gaps remain in the middle of the translational pipeline and approaches are needed to support research at this juncture, where population-based samples with rich environmental and phenotypic measurements can be used to follow-up disease markers identified in GWAS. Specifically, systematic approaches are needed to sift the results of numerous association studies and distill the most promising set of markers for further investigation. These approaches must be able to harness the power of existing resources and to flexibly accommodate new data produced by the fast pace of discovery in genome science.

A key hurdle for research using GWAS results is that risk SNPs identified in GWAS may not cause adverse health outcomes, but may instead be proxies for (correlated with) unmeasured disease-causing variation in the genome.[70, 71] GWAS methods exploit LD across the genome to leverage measurement of 100,000 – 1 million SNPs to capture variation in the 10 million plus SNPs the genome is estimated to contain. The very large sample sizes in GWAS permit detection of risk associations even when proxy SNPs are in imperfect LD with disease-causing variation (correlation<1). GWAS findings are generally applied to smaller samples designed to elucidate etiological and clinical correlates of discovered genes. When GWAS SNPs

are translated to research using smaller samples, the measurement error resulting from imperfect LD with disease causing variants can attenuate associations below levels these samples are powered to detect. Genetic risk scores (GRSs) summarize risk-associated variation across the genome[63] by aggregating information from multiple risk SNPs (the simplest GRSs count disease-associated alleles). Because GRSs pool information from multiple SNPs, each individual SNP is less important to the summary measurement and the "signal" from the GRS is robust to imperfect linkage for any one SNP. For the same reason, GRSs are less sensitive to minor allele frequencies for individual SNPs. As the number of SNPs included in a GRS grows, the distribution of values approaches normality, even when individual risk alleles are relatively uncommon.[72] Therefore, the GRS can be an efficient and effective means of constructing genome-wide risk measurements from GWAS findings.

Obesity is a public health problem that is well suited to risk assessment using a GRS. It is highly prevalent;[73] it is a significant source of health-care costs, morbidity, and mortality;[74-76] it is under strong genetic influence; [13] and GWAS are beginning to elucidate its molecular genetic roots.[77] Therefore, translational research in obesity genomics may ultimately help to address a public health priority. A key challenge is that obesity's genetic roots are diffuse, multifactorial, and non-deterministic; many variants scattered across the genome each contribute small risks for obesity.[78] In other words, information from multiple genetic variants is needed to characterize genetic susceptibility to obesity. Thus, a GRS may be useful. A further challenge is uncertainty about the specific genetic variants to be included in an obesity GRS. Different GWAS identify different genomic loci and, when loci are replicated across GWAS, the specific SNPs identified may be different.[79] To address this challenge, we developed a 3-stage approach to review GWAS results and select specific SNPs to include in a GRS. We devised our approach to be systematic and replicable and to leverage the discovery potential of GWAS while minimizing risk for including false-positive markers. In this article, we describe this 3-stage approach, apply it to develop a GRS for obesity, and test the GRS as a measure of obesity risk using data from the population-based Atherosclerosis Risk in the Communities (ARIC) Study.

METHODS

Sample. The ARIC sample is described elsewhere.[44, 80] Briefly, ARIC is a prospective epidemiologic cohort study sponsored by the National Heart, Lung, and Blood Institute to investigate the etiology of atherosclerotic disease. The study draws from 4 US communities:

Minneapolis MN, Washington County MD, Forsyth County, NC, and Jackson MS. Participants were examined first during 1987-1989, and at 3 subsequent occasions (1990-1992, 1993-1995, and 1996-1998), with ongoing follow-up conducted annually by telephone. ARIC cohort genotype data from the Affymetrix Affy 6.0 Chip and selected phenotypes were obtained for this study from the NIH dbGaP.

The original ARIC sample includes 15,792 participants (27% African American, 55% female). The publicly available dataset obtained from dbGaP for this study includes genotype and phenotype data for 12,771 individuals. Of this sample, 1,212 participants had a missing call rate >2% for SNPs called successfully in ≥95% of the sample and were excluded from subsequent analyses per quality control recommendations of the GENEVA ARIC Project.[81] In addition, although the ARIC study design did not aim to include relatives, genomic analysis by the ARIC investigators revealed familial relationships at the level of half-siblings or closer among 1,674 participants. One member was selected at random from each of the 105 "families" to form a sample of unrelated persons. After these exclusions, the sample consisted of 10,745 participants (23% African American, 55% female, hereafter the "analysis sample").

Body Mass Index and Obesity. Body mass index (BMI: kg/m^2) was calculated from measurements of weight to the nearest pound and height to the nearest centimeter. Obesity was defined according to U.S. Centers for Disease Control and Prevention Criteria as BMI≥30. Anthropometric measurements were collected from participants wearing a scrub suit and no shoes at the 4 in-person data collections.

Genotypes. Details on the genotyping of the ARIC sample are available through dbGaP and are described elsewhere.[82] Briefly, genotyping was conducted by the Broad Institute using the Affymetrix Affy 6.0 SNP array and the Birdseed calling algorithm.[83] Following guidelines for the use of genotypic data provided by the ARIC GWAS team, data were extracted for all SNPs with a sample-wide call rate ≥95%, fewer than 5 discordant calls across duplicated DNA samples in the quality control subsample (n=334), and in Hardy-Weinberg Equilibrium (p>0.001).

Genetic Risk Scores. Current mid-pipeline translational studies use either a "best guess" approach or a "top hits" approach to select genetic markers to include in GRSs. The "best guess" approach selects markers identified in association studies that are located in or near genes with plausible biological relationships to the pathophysiology of a phenotype or that demonstrate

strong and replicable association signals.[84-86] The "top hits" approach selects markers with the strongest association signals in a single GWAS, independent of their biological plausibility.[87, 88] Early studies have illustrated the promise of translational research with GWAS markers, but as the field moves forward, more systematic approaches are needed that can better integrate new information from the latest studies. Neither the top-hits nor the best-guess approach provides a systematic and replicable means of integrating results from multiple GWAS. Meta-analysis can accomplish this, but comprehensive meta-analyses are not always available. Moreover, the top-hits and best-guess approaches do not provide a means to select specific SNPs for follow-up, and this problem is not solved by meta-analysis. The approach of selecting the "lead" SNP at a locus, usually the SNP with the lowest p-value in the largest GWAS, is problematic because different GWAS can report different lead SNPs for the same locus because of differences in GWAS chips, genotyping quality, and data handling and analysis decisions. Thus, an approach is needed that facilitates systematic and replicable SNP selection from results of multiple GWAS.

Our 3-stage approach integrates public-access resources including continuously updated databases of GWAS results, web-based whole-genome analysis tools, and genome-wide data to identify the most promising set of single nucleotide polymorphisms (SNPs) for follow-up. Most importantly, the 3-stage approach addresses key limitations of the top-hits and best-guess approaches: It provides a systematic and replicable means of integrating findings across multiple GWAS and of selecting SNPs for follow-up in new samples. The 3 stages are:

Stage 1) Extraction: All SNPs associated with one of the selected phenotypes at a given significance threshold are "extracted" from each GWAS and retained for further analysis.

Stage 2) Clustering: Extracted SNPs are "clustered" according to patterns of linkage disquelibrium (LD) determined from a reference population that matches the population in the GWAS included in Stage 1. Clustering yields a set of "LD blocks."

Stage 3) Selection: Statistical significance and replication are evaluated at the level of the LD block. The original GWAS results are used to assign a minimum p-value and a replication count for each LD block. The minimum p-value is the lowest p-value reported for any SNP in the LD block in any GWAS contributing data in Stage 1. The replication count is the number of GWAS that reported an association for any SNP in the LD block at the threshold defined in Stage 1.

We applied our 3-stage approach to construct two GRSs for obesity. First, we considered only GWAS published in print or online through December 31, 2008. We chose these GWAS because they were used in previous research that created "top-hits" and "best-guess" obesity GRSs. Thus, we used these GWAS to construct a GRS using our 3-stage approach and compared it to two published GRSs.[61, 62] Second, we considered all GWAS published through December 31, 2010. We applied our 3-stage approach to results from the full set of GWAS and compared the resulting GRS to a top-hits GRS generated from the largest meta-analysis of BMI GWAS published to date[89] and to a best-guess GRS generated from the full set of obesity-associated SNPs reported in the National Human Genome Research Institute (NHGRI) GWAS Catalog.[79] The derivation of the GRS using the 3-stage approach is described in detail in the supplemental material (**Appendix A**). Analyses described in the supplemental material revealed that the 3-stage approach created GRSs that were at least as predictive of BMI and obesity as GRSs created with the top-hits and best-guess approaches. Further analyses to refine the 3-stage approach GRS yielded a final set of 32 SNPs (see Appendix A). We applied 2 weighting schemes to the 32 SNPs before summing them to create our obesity GRS: 1) equal weighting, under which the score was a simple count of BMI-increasing alleles; and 2) effect-size weighting, under which BMI-increasing alleles were weighted by the effect size reported for that locus in the GIANT Consortium [89] or DeCode [90] BMI GWAS. Effect-size weights were adjusted for LD between the SNP tested in the GWAS and the SNP genotyped in the ARIC sample. Each of the 32 SNPs in the GRS was missing for fewer than 1% of participants in any gender/ethnicity cell. GRSs were prorated by dividing the GRS by the number of SNPs contributing data and multiplying by 32. The SNPs included in the final obesity GRS, their BMI-increasing ("effect") alleles, nearby genes, and weights are reported in **Table 2.1**.

Evaluation of the Obesity GRS. Associations between the GRS and obesity-related traits (BMI, weight, waist circumference, obesity) were tested with linear and logistic regression models. These and subsequent models were adjusted for demographic and geographic control variables: age was specified as a linear and a quadratic term; a product term was included for the interaction between age and sex to account for sex differences in BMI and obesity distributions at different ages; the 4 ARIC Study Centers where participants were enrolled in the study were entered as a series of dummy variables (this collection of variables is referred to hereafter and elsewhere in the manuscript as demographics and geography). Predictiveness of the GRS was evaluated using 3 metrics that are established tools for evaluating risk markers in

general [91] as well as for the specific case of genetic risk scores:[92] 1) R^2, the proportion of variation explained in BMI. R^2 was estimated using demographics and geography-adjusted linear regression models. 2) AUC, the area under the receiver operating characteristic curve for obesity, also known as the discrimination index. The AUC corresponds to the probability that a randomly selected obese case will have a higher GRS as compared to a randomly selected non-obese control. A marker that discriminates no better than chance has an AUC of 0.50. A marker that discriminates perfectly has an AUC of 1. A related metric is the partial AUC (PAUC). The PAUC sets a specificity threshold and calculates an AUC-like statistic specific to that specificity. Analyses of PAUC for the GRS set specificity at 80% (the bottom 5th of the ROC curve). AUC and PAUC analyses were stratified by ARIC Study Center using Pepe's method.[93] To determine whether the GRS improved discrimination over and above demographic and geographic information, we calculated a second set of statistics, delta AUC and delta PAUC. Probit regression models were used to generate predicted probabilities of obesity for each ARIC participant using a baseline model that included demographic and geographic information and a test model that also included the GRS. AUCs and were calculated using these predicted probabilities as "risk scores," [94] and estimates of the differences between the baseline and test models were bootstrapped to obtain confidence intervals. AUC analyses were conducted using the Stata package "comproc. [95] 3) IDI, the integrated discrimination index for obesity. The IDI evaluates the added predictiveness of a marker by comparing predictions made using a baseline set of risk markers to predictions that also include information about the new risk marker:

$$IDI=(Prob_{test,\ obese} - Prob_{test,non\text{-}obese}) - (Prob_{baseline,\ obese} - Prob_{baseline,\ non\text{-}obese})$$

where "Prob" is the average predicted probability for a particular group from a particular model. The IDI measures change in model sensitivity net of change in model specificity and is a more sensitive measure than delta AUC.[96] An IDI of zero indicates that the test model performs comparably to the baseline model. Positive IDI values index net improvement in model sensitivity. Baseline and test models for IDI analyses were identical to those used in delta AUC analyses.

We tested differences between the predictiveness metrics for different risk scores by bootstrapping confidence intervals around the R^2 and AUC metrics (comparing the difference in estimated metric values across 1,000 random samples drawn with replacement from the ARIC database[95]) and by applying Pencina's method [96] to test change in the IDI metric. Comparisons

were as follows: Un-Weighted GRS vs. Weighted GRS; Weighted GRS vs. Simple Genetic Risk Assessment (the sum of risk alleles at the two best-replicated obesity loci, in the gene *FTO* and downstream of the gene *MC4R*, rs9939609 and rs12970134, respectively); Weighted GRS vs. Socioeconomic Index (Educational attainment measured in 5 categories: grade-school or less, some high school, high school graduate, vocational school, college, graduate/professional school, **Appendix A**, **Supplementary Table 2.8**).

RESULTS

Obesity risk-allele distributions were similar for males and females, but were different for whites and African Americans. The variance of the un-weighted GRS was greater for whites as compared to African Americans (SD= 3.50 as compared to 3.25, $p<0.001$ using Brown and Forsythe's method [97]), as was the mean (M=28.80 as compared to 24.87, $p<0.001$ using t-test for unequal variances; see also **Appendix A, Supplementary Figure 2.1**). This difference reflected lower frequencies of BMI-increasing alleles for several GRS SNPs among African American ARIC participants (**Table 2.1)**. Subsequent analyses were stratified by race.

The obesity GRSs were weakly but consistently associated with BMI and the probability of being obese among whites and African Americans, but associations were weaker among African Americans (**Figure 2.1**). Among whites, after adjusting for age, sex, and geography, the un-weighted GRS was associated with BMI at r=0.12 and the weighted GRS was associated with BMI at r=0.13 ($p<1x10^{-26}$ for both). This effect size corresponded to a 0.60-unit increase in BMI per 1-standard-deviation increase in the GRS. For each 1-standard-deviation increase in their un-weighted and weighted GRSs, a white ARIC participant's risk for obesity increased by 19.35% and 20.51%, respectively ($p<1x10^{-18}$ for both). Among African Americans, the weighted and un-weighted GRSs were associated with BMI at r=0.05 ($p<0.05$ for both). For each standard deviation increase in their un-weighted and weighted GRSs, an African American ARIC participant's risk for obesity increased by 3.54% (p=0.059) and 4.92% (p=0.017), respectively. Results were substantively unchanged when control variables were removed from the models. We conducted a series of additional sensitivity analyses to evaluate heterogeneity in GRS associations (described in detail in **Appendix A**). These analyses supported a linear association between the GRS and BMI; showed that GRS-BMI associations were similar to GRS-weight and GRS-waist circumference associations; and revealed no sex or age differences in GRS-BMI associations.

The obesity GRSs performed similarly on the 3 predictiveness metrics (**Table 2.2**). The top panel of Table 2 addresses clinical validity. It presents the 3 metrics for the un-weighted and weighted GRSs. Among whites, weighted and un-weighted obesity GRSs explained small, but statistically significant proportions of the variance in BMI (R^2), discriminated obese from non-obese participants modestly better than chance (AUC), and contributed small net improvements to the sensitivity of an obesity prediction model over and above demographic and geographic information (IDI). Among African Americans, the GRS did not contribute to the explanation of variance in BMI over and above demographic and geographic information, to the discrimination of obese from non-obese participants, or to the net sensitivity of the obesity prediction model. Use of weights derived from BMI GWAS improved the performance of the GRS among whites and African Americans, but this improvement was not statistically significant ($p>0.10$ for all comparisons).

The bottom panel of Table 2 addresses research utility. It presents predictiveness metrics for two comparison measures of obesity risk: the simple genetic risk assessment (weighted combinations of rs9939609 in *FTO* and rs12970134 downstream of *MC4R*) and the socioeconomic index (a 5-category measure of educational attainment). The *FTO* and *MC4R* loci and socioeconomic status are robust correlates of BMI and obesity in adult samples.[98, 99] Comparison of the 32-locus GRS to a two-locus risk assessment can illustrate whether the GRS offers value added over a simpler genetic risk assessment. Comparison of the GRS to socioeconomic status can illustrate how the predictiveness of the GRS compares to the predictiveness of a social determinant of obesity that is not easily changed but that is understood to be important in etiological research.[100] Among whites, the genetic risk scores performed better than the comparison measures of obesity risk on all 3 metrics ($p<0.01$ for all comparisons). Among African Americans, the GRSs performed no differently from the simple genetic risk assessment ($p>0.10$) and performed less well as compared to the socioeconomic index ($p=0.021$). When combined with the comparison risk measures and with demographic and geographic information, the GRS improved predictiveness for whites but not for African Americans (**Appendix A, Supplementary Table 2.9**).

Figure 2 shows the model-based receiver operating characteristic curves for a baseline model that included demographic and geographic information and a test model that also included the weighted GRS. The change in AUC from the baseline model to the test model was

greater than zero (Delta AUC=0.048, 95% CI 0.313-0.658, $p<10^{-7}$), indicating that the GRS improved discrimination of obese cases. This improvement in discrimination was concentrated at low specificities, but extended to the portion of the ROC curve of greatest interest to clinicians. At a specificity of 0.8, the test model including the GRS was marginally more sensitive as compared to the baseline model (Delta Partial AUC=0.007, 95% CI <0.0003-0.010, p=<0.001). Results for African Americans are presented in **Supplementary Figure 2.2 (Appendix A)**.

As a final analysis, we asked whether the obesity GRS was associated with mortality risk. The ARIC study conducted follow-up with participants through December 31, 2004 to determine whether study members had died. Mortality follow-up data were available for 8,284 of the 8,286 white participants in our analysis sample. 15% of this sample (n=1,253 individuals) died during the 17 years of follow-up from the first study visit. We analyzed mortality risk using Cox proportional hazard models to adjust for demographic and geographic factors. Independent of demographics and geography, individuals with higher genetic risk scores were more likely to die during the follow up period (Hazard Ratio=1.12, 95% CI [1.04-1.15]). Consistent with analyses of BMI and obesity, the GRS was not associated with mortality among African Americans. **Figure 2.3** presents cumulative mortality hazards for white ARIC participants in the top, middle, and bottom quintiles of the genetic risk distribution. The mortality hazard associated with the GRS did not depend on individuals' BMIs. Adjustment of the mortality hazard model for BMI only slightly reduced the mortality hazard associated with genetic risk (Hazard Ratio=1.10 [1.04-1.17]).

DISCUSSION

We used a 3-stage approach to construct an obesity GRS from GWAS results. Our tests of this obesity GRS in the population-based ARIC cohort revealed it to be a highly statistically significant predictor of BMI measured at 4 time points across 10 years, of weight and waist-circumference, and of obesity. In terms of value added, the GRS improved prediction of BMI and obesity over and above demographic and geographic information, *FTO* and *MC4R* genotypes, and information about socioeconomic status. Thus, the GRS provides a measure of genetic predisposition to obesity that could inform etiological and treatment research. Finally, the GRS

was associated with mortality risk. Interestingly, higher mortality risk for individuals with higher GRSs did not depend on their BMI.

The research utility of the GRS is likely limited to samples of European descent. GRS-BMI and GRS-obesity associations in African American ARIC participants were much smaller than comparable associations in white ARIC participants. Although the sample included fewer African Americans than whites, power to detect effects of equal size to those observed in whites was well over 80% in the African American sample. Moreover, effect-size measures (r, R^2, relative risk, AUC, IDI) showed little evidence that the GRS predicted BMI or obesity among African Americans. These results suggest caution in using GWAS of European-descent populations to derive GRSs for African Americans. Our analyses indicated the GRS performed similarly among men and women. However, emerging evidence for gene-sex interactions in obesity [101, 102] suggests that future obesity GRSs may require sex-specific construction.

Our results have implications for theory, research, and clinical practice. With respect to theory, our results are consistent with the hypothesis that genetic risk for obesity is quantitatively distributed and can be operationalized in a GRS.[65] With respect to research methods, our findings illustrate one approach to operationalize quantitative genetic risk. A systematic and replicable approach to selecting SNPs from association studies to follow-up in etiological and treatment research will be especially important with the advent of next-generation sequencing approaches. Next generation sequencing is likely to uncover many new disease-associated loci for obesity and for other phenotypes of interest to clinicians and researchers. These variants, though rarer in the population, may have higher penetrance and thus greater clinical relevance. Future research can also make use of the GRS derived in this study as a measure of inherited obesity risk. With respect to clinical practice, results indicate that, for persons in middle age, GWAS SNP-based approaches to obesity risk assessment offer little in the absence of more detailed information about lifestyle and environment. Although genetic information reliably predicted risk for obesity over and above demographics and geography, the magnitude of this additional risk was insufficient to recommend our score for use in clinical risk assessments. This result is especially important in the context of questions over consumer genomics services.[103] Our 3-stage approach derived a more comprehensive genetic risk assessment for obesity than those currently used by companies marketing genomics services directly to consumers. The very modest risk information furnished by our GRS

recommends caution on the part of consumers and health professionals in interpreting risk information provided by consumer genomics companies. The standard of evidence used here—multi-method assessment of predictiveness in large, population-based samples--should be considered a minimum standard for the validity of such risk information.

These results should be considered in light of the following limitations: First, some ARIC participants were included in the samples of some of the GWAS used to construct the GRS. However, these ARIC participants represented a minority of the GWAS samples and results in the ARIC sample are similar to results from samples not included in any of the GWAS.[61, 62] Second, some risk loci identified by our 3-stage approach could be genotyped in the ARIC sample using only relatively weak proxies. Given the small improvement to predictiveness associated with each additional SNP included in the GRS, it is unlikely that this limitation influenced the substance of our results, but it is possible that our GRS is moderately more predictive than analyses in the ARIC cohort suggest. Third, our analyses were limited to African American and white Americans. The ARIC cohort does not contain Asian-descent or Hispanic individuals. It remains unclear whether the relatively greater similarity between these and European populations [104] would support the generalization of our GRS. However, GWAS of Asian and Hispanic samples [88, 105] suggest that a European-descent population-derived GRS may omit important risk loci for these populations. As more GWAS of non-European populations become available, our 3-stage approach can be used to derive additional population-specific GRSs. Fourth, there is mounting evidence that many genetic factors predisposing individuals to obesity are sex specific [106] and that GWAS that fail to model such sex specificity may not detect important risk variants.[107] Results from GWAS modeling gene-by-sex interaction support this hypothesis.[101, 102, 108] As more such GWAS become available, our 3-stage approach can be used to derive sex-specific GRSs for obesity. Finally, the ARIC sample is limited to individuals in middle age. There is evidence that genetic risk for obesity has dynamic consequences across development.[109, 110] It will be important in subsequent investigations to evaluate our obesity GRS in longitudinal cohorts that capture a broader section of the life course, and particularly in young people, as they are a key prevention target.[10]

We constructed a GRS for obesity and showed that it predicted BMI and obesity in a population-based sample of middle-aged adults. We further showed that this GRS was longitudinally associated with mortality risk. These associations suggest that future research into

obesity etiology and treatment can make use of genetic information. However, our analyses do not support the use of genetic testing for individual-level obesity-risk prediction. Future research with this GRS should characterize the expression of genetic risk across the life course and particularly during childhood, when intervention to prevent the development of obesity may be most effective.

Table 2.1. Single nucleotide polymorphisms included in the obesity genetic risk score (GRS). Alleles are reported from the forward strand. The GRS was computed by counting the number of effect alleles at each SNP, multiplying that number by the SNP's weight, and then summing the results across the set of 32 SNPs. Weights reflect per-allele changes in BMI estimated in the the GIANT Consortium GWAS meta-analysis [89] except for rs867559, for which the weight was estimated in the DeCODE GWAS meta-analysis [90].

						Effect Alle Frequency (ARIC Sample)	
Chr	Nearby Gene	SNP	Effect Allele	Other Allele	Weight	Whites	African Americans
1	NEGR1	rs2815752	G	A	0.13	62%	55%
	TNNI3K	rs1514175	A	G	0.07	43%	68%
	PTBP2	rs1555543	A	C	0.06	58%	43%
	SEC16B	rs543874	G	A	0.22	20%	25%
2	FANCL	rs759250	A	G	0.10	29%	8%
	LRP1B	rs2121279	T	C	0.08	14%	3%
	TMEM18	rs2867123	G	C	0.30	83%	88%
	RBJ	rs10182181	G	A	0.14	54%	16%
3	CADM2	rs12714640	A	C	0.10	19%	6%
	ETV5/DGKG	rs1516728	T	A	0.11	77%	48%
4	GNPDA2	rs12641981	T	C	0.18	43%	23%
	SLC39A8	rs13114738	T	C	0.13	8%	1%
5	POC5 FLJ35779	rs10057967	C	T	0.10	63%	51%
	ZNF608	rs6864049	A	G	0.07	54%	81%
6	TFAP2B	rs734597	A	G	0.13	17%	9%
9	LING02 LRRN6C	rs1412235	C	G	0.11	31%	16%
	LMX1B	rs867559	G	A	0.24	20%	32%
11	RPL27A	rs2028882	C	A	0.06	50%	34%
	BDNF	rs10501087	C	T	0.18	79%	93%
	MTCH2	rs12419692	A	C	0.05	36%	9%
12	BDCDIN3D, FAIM2	rs7138803	A	G	0.12	38%	17%
13	MTIF3, GRF3A	rs1475219	C	T	0.09	21%	22%
14	PRKD1	rs1440983	A	G	0.15	5%	23%
	NRXN3	rs7144011	T	G	0.13	22%	24%
15	MAP2K5	rs28670272	G	A	0.13	77%	59%
16	GPR5B	rs11639988	G	A	0.17	85%	76%
	ATXN2L, TUFM, SH2B1	rs12443881	T	C	0.15	39%	9%
	FTO	rs9939609	A	T	0.38	41%	48%
18	MC4R	rs12970134	A	G	0.21	26%	13%
19	KCTD15	rs11084753	A	G	0.04	67%	64%
	QPCTL	rs11083779	C	T	0.07	96%	89%
	ZC3H4 TMEM160	rs7250850	G	C	0.09	71%	20%

Table 2.2. Predictiveness Metrics for the 3-Stage Approach Obesity Genetic Risk Score and Comparison Measures of Risk for Obesity. The simple genetic risk score is a component of the weighted obesity genetic risk scores. Values of R^2 were estimated using linear regression models and reflect the improvement in the proportion of variance explained by the model beyond the baseline prediction derived from demographic and geographic information. Percentile-based confidence intervals (CIs) were generated using the bootstrap method. Areas Under the Curve (AUCs) and CIs were estimated from ROC curves constructed for raw values (i.e. actual values of the measures tested rather than predicted values generated from a regression model) and were adjusted for the ARIC Study Center where data were collected. Integrated Discrimination Indexes (IDIs) and test statistics were estimated for comparisons of a baseline model including demographic and geographic information to a test model that included this information and the GRS.

	White ARIC Participants (n=8,286)			Black ARIC Participants (n=2,442)		
	R^2 (95% CI)	AUC (95% CI)	IDI (p-value)	R^2 (95% CI)	AUC (95% CI)	IDI (p-value)
Panel A. Predictivness of the un-weighted and weighted obesity GRSs						
Un-Weighted GRS	1.39% (0.94% - 1.89%)	0.565 (0.550 - 0.581)	0.009 (4.65E-18)	%0.11 (-%0.04 - %0.57)	0.515 (0.491 - 0.540)	0.001 (0.067)
Weighted GRS	1.57% (1.11% - 2.10%)	0.570 (0.554 - 0.584)	0.010 (8.25E-20)	%0.14 (-%0.03 - %0.65)	0.521 (0.497 - 0.544)	0.002 (0.152)
Panel B. Predictiveness of comparison risk measures						
Simple Genetic Risk Assessment: FTO & MC4R-linked SNPs only	0.59% (0.31% - 0.97%)	0.543 (0.528 - 0.557)	0.004 (3.54E-09)	-%0.02 (-%0.04 - %0.25)	0.516 (0.493 - 0.539)	0.001 (0.149)
Socioeconomic Index: 5-category measure of educational attainment	0.57% (0.29% - 0.87%)	0.532 (0.517 - 0.546)	0.003 (7.83E-07)	%1.06 (%0.42 - %1.99)	0.561 (0.538 - 0.584)	0.016 (2.71E-11)
Panel C. Predictiveness of model-based risk assessments (including demographic and geographic information)						
Simple Genetic Risk Assessment	3.88%	0.550		5.35%	0.607	
Weighted GRS	4.88%	0.574		5.52%	0.609	
Change in predictiveness with addition of weighted GRS to model	1.00% (0.58%-1.42%)	0.024 (0.012-0.036)	0.006 (7.81E-13)	0.17% (-%0.15-%0.51)	0.002 (-0.005-0.009)	0.001 (0.055)
Socioeconomic Status	4.70%	0.550		7.70%	0.643	
Socioeconomic Status + weighted GRS	6.20%	0.586		7.92%	0.645	
Change in predictiveness with addition of weighted GRS to model	1.50% (1.00%-1.99%)	0.036 (0.023-0.050)	0.010 (5.46E-19)	0.22% (-%0.14-%0.55)	0.002 (-0.003-0.008)	0.002 (0.012)

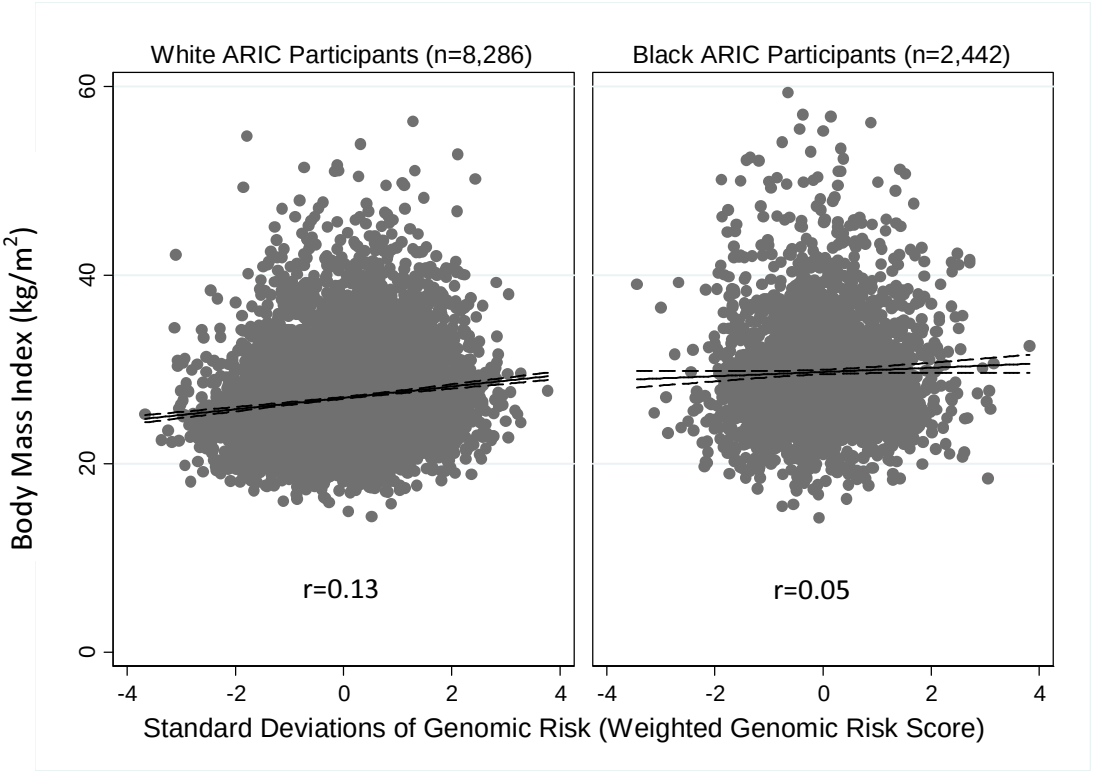


Figure 2.1 Panel A. BMI for White and African American ARIC Participants Plotted Against the Weighted Obesity Genetic Risk Score. Dashed outlines represent 95% confidence intervals. Pearson correlations (r) were adjusted for gender, age and ARIC Study Center where data were collected. Removal of outliers (not shown) did not alter correlation estimates at the third decimal point. Correlations were statistically significant for white ($p<1x10^{-30}$) and African American (p=0.014) ARIC participants.

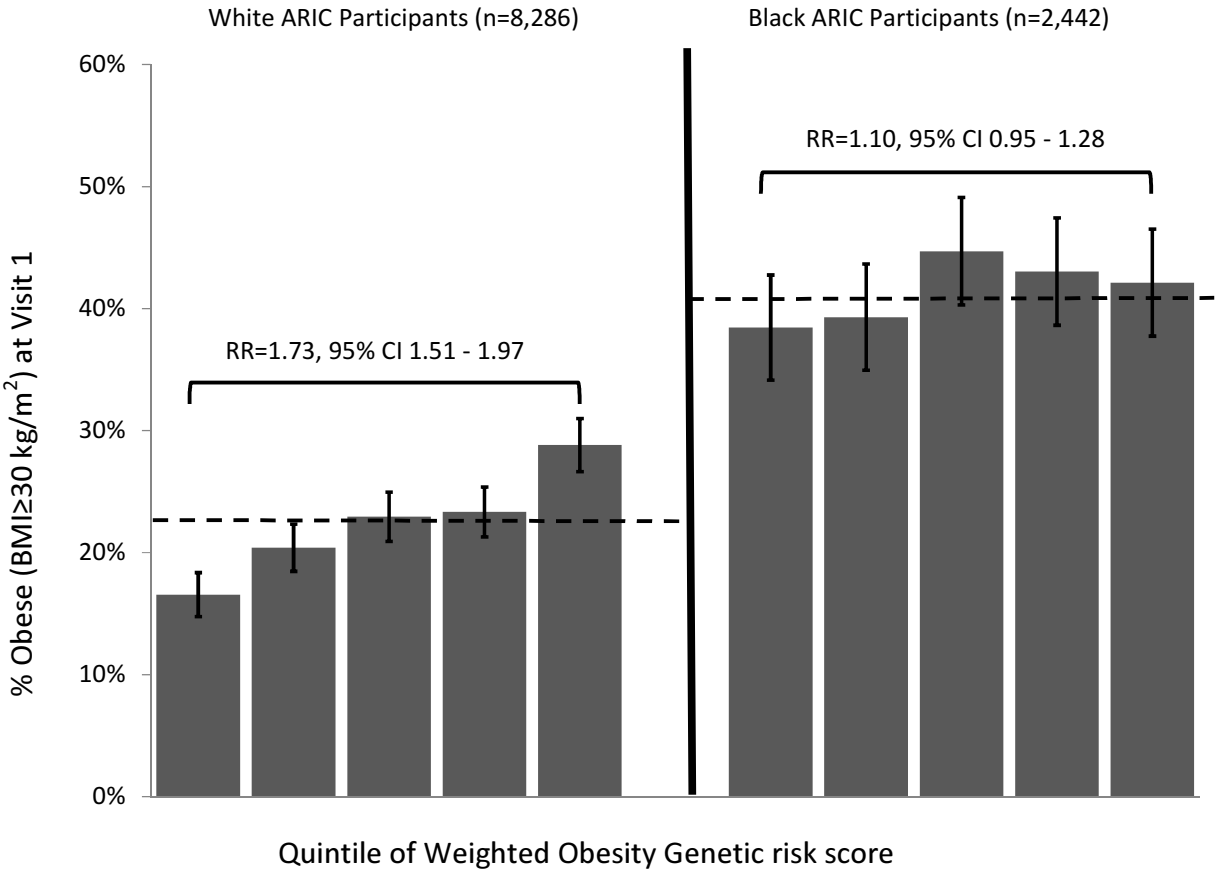


Figure 2.1 Panel B. Percentage White and African American ARIC Participants Who Were Obese (BMI≥30kg/m^2) at the First Study Visit, by Quintile of Genetic Risk Score. Quintiles were determined separately for whites and African Americans. Error bars represent 95% confidence intervals. Risk ratios are for comparisons of highest to lowest quintiles of genomic risk and were estimated with adjustment for gender, age, and ARIC study center where data were collected. Dashed lines represent sample means. Among white ARIC participants, all quintile to quintile differences are statistically significant ($p<0.01$), with the exception of the 3rd and 4th quintiles. Among African American ARIC participants, the percent obese in the lowest quintile was lower than in the third and fourth quintiles ($p<0.05$).

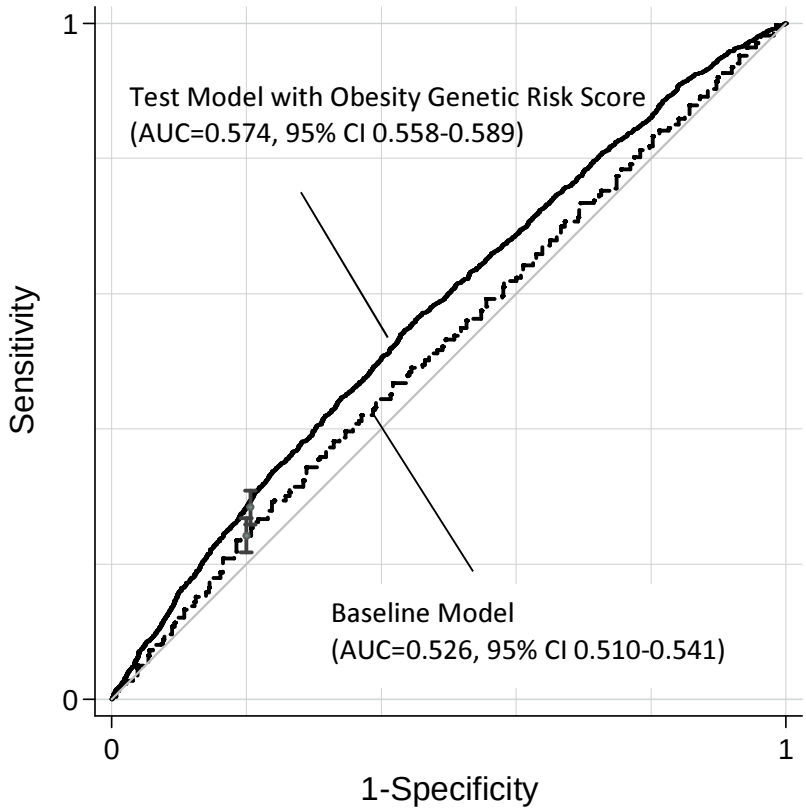


Figure 2.2. Receiver Operating Characteristic Curves for Obesity Among White ARIC Participants (n=8,286). Baseline Model = gender, age (quadratic), gender x age interaction, ARIC study center; Test Model = baseline model + weighted obesity genetic risk score. ROC curves were constructed using predicted values from probit regressions of obesity (BMI≥30) on the model terms. Delta AUC (AUC_{Test}-$AUC_{Baseline}$) = 0.048, 95% CI 0.031-0.066, $p<1x10^{-7}$. Delta Partial AUC at 80% specificity=0.007, 95% CI 0.003-0.010, p<0.001. AUCs, partial AUCs, and delta AUCs were estimated using Pepe's method [93, 95].

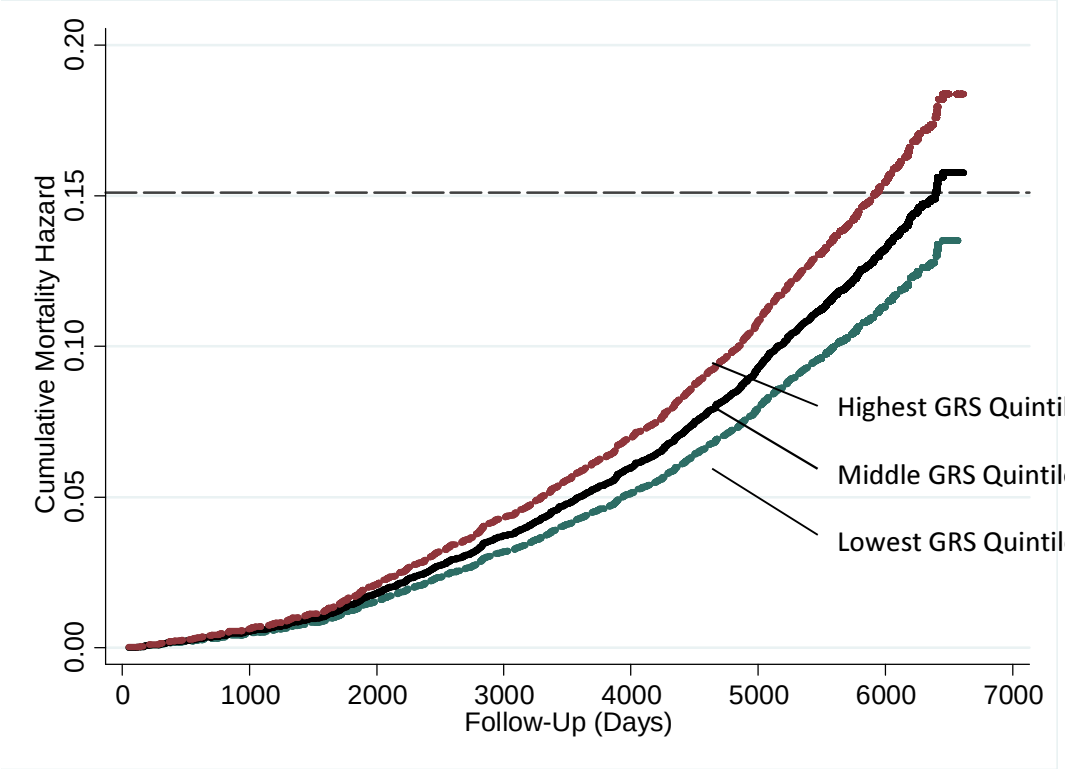


Figure 2.3. Cumulative Mortality Hazards for White ARIC Participants in the Highest, Middle, and Lowest Genetic Risk Score (GRS) Qunitiles. Hazards were estimated from a Cox proportional hazard model adjusted for age, sex, the age-sex interaction, and the ARIC Study Center where data were collected. The dashed line represents sample-wide mortality at the end of follow-up (15%). By the end of follow-up, unadjusted mortality was 12.17% in the lowest GRS quintile, 15.48% in the middle GRS quintile, and 17.32% in the highest GRS quintile.

CHAPTER 3.

POLYGENIC RISK FOR ADULT OBESITY IS MEDIATED BY RAPID CHILDHOOD GROWTH: EVIDENCE FROM A 4-DECADE LONGITUDINAL STUDY[3]

INTRODUCTION

Obesity is known to be heritable and genome-wide association studies (GWAS) have begun to uncover the molecular roots of this "heritability" by identifying multiple single-nucleotide polymorphisms (SNPs) associated with higher adult body mass index (BMI).[78] The next step is to understand how these SNPs influence the development of obesity. Individual differences in obesity risk emerge during gestation and are further established during infancy and childhood through accelerated growth trajectories.[46, 132] Therefore, examination of developmental phenotypes in relation to genetic risk represents a promising approach to understand the pathogenesis of obesity.[109, 110, 133] In this study, we asked how SNPs with replicated GWAS evidence for association with adult BMI relate to growth across the first four decades of life and to adult obesity in a birth cohort followed prospectively from birth through age 38 years.

SNPs identified in GWAS contribute small increments to obesity risk.[49] Aggregating GWAS-identified SNPs to produce a genome-wide index (a "genetic risk score") yields a quantitative measure of inherited predisposition towards a trait, such as BMI.[65] This approach has shown promise in the study of complex diseases such as diabetes and heart disease.[55, 56] In this study, we used a multi-locus genetic-risk score to test how a genetic predisposition to higher adult BMI might also relate to developmental phenotypes of growth during proposed critical periods in the development of obesity. Three developmental phenotypes are of interest: Growth during gestation, postnatal growth, and the adiposity rebound. All correlate with adult BMI and are thought to program risk for adult obesity.[10, 134, 135] Therefore, we tested the hypothesis that polygenic risk for adult obesity is mediated by these developmental phenotypes of rapid early

[3] In press, *Archives of Pediatrics & Adolescent Medicine*; with TE Moffitt, RM Houts, GG Bennett, AK Biddle, JA Blumenthal, JP Evans, HL Harrington, K Sugden, B Williams, R Poulton, and A Caspi

growth (**Figure 1**). Understanding when in development genetic risk for obesity is expressed can help to refine research and intervention targets.

If genetic risk is mediated through early growth, it would be important to know how measured genetic risk compares to parental BMI in predicting children's growth and obesity risk. We thus tested whether obesity risk information contained in the genetic risk score was independent of obesity risk information contained in the BMIs of children's parents. That is, does the genetic risk score contain novel information about children's risk for obesity over and above their family history?

METHODS

PARTICIPANTS

Participants are members of the Dunedin Multidisciplinary Health and Development Study, a longitudinal investigation of health and behavior in a complete birth cohort. Study members (N=1,037; 91% of eligible births; 52% male) were all individuals born between April 1972 and March 1973 in Dunedin, New Zealand, who were eligible for the longitudinal study based on residence in the province and who participated in the first follow-up assessment at age 3. The cohort represents the full range of socioeconomic status in the general population of New Zealand's South Island and is primarily white. Assessments were carried out at birth and at ages 3, 5, 7, 9, 11, 13, 15, 18, 21, 26, 32, and, most recently, 38 years, with over 90% retention. At each assessment, study members are brought to the Dunedin research unit for a full day of interviews and examinations. The Otago Ethics Committee approved each phase of the study. Informed consent was obtained from all study members.

MAIN EXPOSURES

Obesity Genetic Risk Score. We derived a 32-SNP genetic risk score (GRS) from published genome-wide association studies (GWAS) of body mass index (BMI), obesity, weight, and waist circumference in European-descent populations. The construction of the GRS is described in the **Appendix B**. We validated our GRS as a measure of obesity risk in data from the Atherosclerosis Risk in the Communities (ARIC) sample.[81] European-descent individuals in the ARIC sample with higher GRSs were larger as measured by BMI, weight, and waist circumference

($r>0.10$, $p<1x10^{-20}$), and were more likely to be obese (Relative Risk (RR)=1.73, 95% CI 1.51-1.97 for individuals in the highest vs. the lowest quintile of the GRS distribution).

We genotyped the 32 GRS SNPs in the Dunedin Study cohort with the Illumina BeadPlex Array using DNA extracted from whole blood (93% of the sample) or buccal swabs (7% of the sample). Of the 32 GRS SNPs, 29 were called successfully in >95% of the cohort and we constructed the final score from these SNPs (**Appendix B, Supplemental Table 3.1**). Comparison of the 29-SNP GRS to the original 32-SNP GRS in the ARIC sample revealed no differences in score distribution or effect sizes. Dunedin Study members carried between 15 and 36 risk alleles (Mean (M)=26.04, Standard Deviation (SD)=3.32). After weighting, GRS values ranged from 13.71-35.04, M=24.71, SD=3.59 (**Appendix B, Supplemental Figure 3.1**). The GRS was standardized to have a mean of 0 and a standard deviation of 1 for analyses.

Family History of Obesity. Parent BMI was available for 98% of the cohort. Parents' BMIs were computed from self-reports of height and weight when children were aged 11 years. To measure familial predisposition to obesity, BMIs of parents were standardized within sex and the standardized scores were averaged to create a single family-history score.

OUTCOME MEASURES

BMI. Individuals' height and weight were measured at ages 3, 5, 7, 9, 11, 13, 15, 18, 21, 26, 32, and 38 years. Height was measured to the nearest millimeter using a portable Harpenden Stadiometer (Holtain, Crymych, UK). Weight was recorded to the nearest 0.1kg using a Lindell Beam Balance at ages 3, 5, 7, 9, 11, 13, 15, and 21 years and calibrated scales at ages 26, 32, and 38 years. Individuals were weighed in light clothing. BMI was computed as weight (kg)/height (m^2). Obesity was defined at age 15 using U.S. Centers for Disease Control and Prevention cut points (BMI≥24.64 for boys, BMI≥25.46 for girls), which show similar predictive validity for obesity and coronary heart disease in young adulthood to International Obesity Task Force cut points.[136] Obesity was defined at ages 18-38 years as BMI≥30. Individuals who met obesity criteria at ≥50% of 6 measurements between 15 and 38 years were classified as chronically obese.[137]

Additional Measures of Adiposity. At ages 7 and 9 years, tricep and subscapular skinfold thicknesses were measured by trained anthropometrists. At ages 26, 32, and 38 years, waist

girth was measured by averaging two measurements of the perimeter at the level of the noticeable waist narrowing. At ages 32 and 38 years, fat mass was measured using the Tanita Body Composition Analyzer BC1418 to assess bioelectrical impedence.[138]

Developmental Phenotypes of Early Growth. Rate of early-childhood weight gain was assessed as the difference between weight at birth (from hospital records) and weight at age 3 years. Adiposity rebound was calculated as the nadir of each individual's childhood BMI curve fitted over ages 3-13 years. We used multilevel longitudinal modeling to fit individual growth curves.[139] Models included linear and quadratic slope terms and were adjusted for sex. Children in our sample experienced adiposity rebound around age 6 years (M=6.11 years, SD=1.10 years) at a BMI around 16 (M=15.57, SD=1.00).

ANALYSES

We analyzed life-course growth using a multilevel longitudinal growth model[139] fitted to BMI measurements at ages 3, 5, 7, 9, 11, 13, 15, 18, 32, and 38 years. We set the intercept at age 13 years. We modeled separate linear and quadratic slopes for growth during childhood (ages 3-13 years) and adulthood (ages 13-38 years). The intercept captured the sample mean BMI at age 13 years (β=19.97). Slope coefficients captured annual change/acceleration in BMI. Linear slope terms captured change in BMI across childhood (β=1.19) and adulthood (β=0.51). Quadratic slope terms captured acceleration of change--the concavity of the trajectory in childhood (β=0.08) and the convexity of the trajectory in adulthood (β=-0.01). All model parameters were statistically significant ($p<0.001$).

We tested genetic influence on growth by modeling the intercept and linear slope parameters of the life-course growth curve as functions of the GRS and covariates. GRS coefficients measured the effect of a one standard deviation increase in genetic risk on BMI at age 13 years (intercept) and on the linear change per-year in BMI between ages 3 and 13 years (childhood slope) and ages 13 and 38 years (adulthood slope).

We tested genetic associations with cross-sectional measurements of BMI and with other quantitative traits using linear regression models. GRS coefficients were standardized to effect-size correlations (Pearson's r) for ease of interpretation. We tested genetic associations with obesity risk using Poisson regression models. GRS coefficients were exponentiated to compute

relative risks (RR). We tested mediation of genetic risk for obesity through developmental phenotypes of early growth using the structural equation described by MacKinnon & Dwyer.[140] Mediation analyses decomposed GRS-obesity associations into direct (un-mediated) and indirect (mediated through a developmental phenotype) components. Statistical tests of mediation were conducted using methods described by Preacher and colleagues.[141-143]

All models were adjusted for sex and comprised the 98% (n=856) of European-descent Dunedin Study members with available body mass index, family history, and genotype data. We used SAS 9.2[144] for growth modeling and mediation analyses and Stata 11.0[145] for other analyses.

RESULTS

Children with higher genetic risk scores (GRSs) were larger and grew faster during childhood and during adulthood. Children with higher GRSs had higher BMIs at every age assessed, from age 3 to 38 years (**Table 3.1**). In the life-course growth model, higher GRSs predicted higher mean levels of BMI (intercept β=0.38, p<0.001), faster growth in childhood (β=0.03, p<0.001) and faster growth in adulthood (β=0.02, p=0.017). **Figure 3.2** shows life-course growth curves for children with high, low, and average GRSs.

To rule out the possibility that variation at the *FTO* locus accounted for our observed GRS-growth associations, we repeated the analysis, adjusting slope and intercept estimates for the *FTO* SNP rs9939609, which is the best replicated GWAS result for BMI,[79] has been shown to influence growth,[98, 109] and carried the largest weight of any SNP in our GRS. GRS-growth associations were unchanged by adjustment for rs9939609. Independent of their rs9939609 genotype, children with higher GRSs were larger across four decades of follow-up (intercept β=0.40, p<0.001) and grew faster during childhood and during adulthood (childhood linear slope β=0.03, p=0.003; adult linear slope β=0.02, p=0.013).

To rule out the possibility that GRS-growth associations reflected associations with height or with muscle mass and not with adiposity, we tested associations between the GRS and childhood skinfold thicknesses and adult waist-girth and fat-mass measurements. These

measurements are less susceptible to inflation as a result of body-size and are considered to be more direct measures of body fat.[138] GRS correlations with these alternative measures of adiposity were statistically significant and were similar to GRS correlations with BMI (**Table 1**).

Children with higher genetic risk scores were at greater risk for obesity across two decades of adult follow-up. As teenagers (ages 15-18 years), 6% of Dunedin Study children had BMIs in the obese range; in their 20s, 11% met criteria for obesity; by age 38 years, 23% met criteria for obesity, consistent with nationwide prevalence among European-descent New Zealanders (http://socialreport.msd.govt.nz/health/obesity.html). 9% of the sample were classified as chronically obese. **Figure 3.3** shows obesity prevalences for children at low (below average) and high (above average) genetic risk. Children at high genetic risk were between 1.61- and 2.41-times more likely to be obese in their teens, 20s, and 30s, and were 1.90-times likely to be chronically obese across 3+ assessments as compared to children at low genetic risk.

Polygenic risk for adult obesity is mediated by developmental phenotypes of rapid childhood growth. To determine whether genetic risk for obesity was mediated through rapid early growth, we investigated relationships among children's GRSs, their growth during gestation and childhood, and their obesity outcomes across two decades of adult follow-up.

The first developmental period theorized to entrain adult obesity risk is gestation. However, the GRS was not associated with fetal growth as indexed by birthweight (r=0.00, p>0.90, **Table 1**). Nevertheless, by age 3 years, children at higher genetic risk had higher BMIs relative to their peers (r=0.08, p=0.043), raising the question of whether growth between birth and age 3 years mediated genetic risk for obesity. Children at higher genetic risk did gain more weight between birth and age 3 years (r=0.09, p=0.014, **Table 3.1**). Consistent with previous research,[146, 147] children with more rapid birth-3 weight gain were more likely to become obese (**Table 3.2**). Decomposition of GRS-obesity associations into direct effects and indirect effects indicated that birth-3 weight-gain mediated statistically significant portions of genetic risk for obesity in the teens and for chronic obesity, but not for obesity in the 20s or 30s individually (**Table 3.2**).

Adiposity rebound, when children begin to gain body fat after losing it during early childhood, is a third period in development theorized to entrain adult obesity. For children at higher genetic risk, adiposity rebound occurred earlier in development and at higher BMI (r=-0.13 for age and r=0.17 for BMI, p<0.001 for both, **Table 3.1**). Consistent with previous research,[134, 148] children with earlier adiposity rebound and higher BMI at adiposity rebound

were more likely to become obese (**Table 3.2**). Decomposition of GRS-obesity associations into direct effects and indirect effects revealed that adiposity rebound mediated large and statistically significant portions of genetic risk for obesity in the teens, 20s, and 30s and for chronic obesity (**Table 3.2**).

The genetic risk score contained information about children's growth and their risk for obesity in adulthood that was not available in their family histories. Higher genetic risk predicted faster growth and increased risk for obesity in children with normal-weight parents and in children with overweight parents (**Figure 3.4 Panels A and B**). That is, the GRS contributed independent and additive information to the prediction of children's growth and their risk for obesity in adulthood over and above family history information (**Appendix B, Supplementary Table 3.2**).

DISCUSSION

We conducted a developmental genetic investigation into the etiology of obesity in a four-decade long prospective birth-cohort study. We measured polygenic risk for obesity using a multi-locus genetic risk score derived from GWAS of obesity-related phenotypes. Our analyses revealed that polygenic risk for obesity was partly mediated by rapid growth in the early childhood years following birth. This finding supported our hypothesis that developmental phenotypes were critical in linking a genetic predisposition to adult obesity. Furthermore, risk for obesity measured by the genetic risk score was independent of risk information available in parental BMI.

These findings have implications for clinical practice and for developmental and epidemiologic research. First, the results suggest promise for utilizing genetic information in obesity risk assessments. Parent BMI has been proposed as a screening measure to target obesity prevention in children on the basis of effect-size correlations only slightly larger than those we report for our genetic risk score.[149] New developments in genome science, including next-generation sequencing, may uncover new variants that further improve the performance of a SNP-based risk assessment.[150-152] Moreover, the genetic risk score contained information about children's future obesity risk that could not be derived from measurements of parents, suggesting that positive family history may not always be an appropriate prerequisite for genetic

testing. Second, our findings illustrate how polygenic influences on development can be investigated using genetic risk scores. Prospective-cohort studies containing repeated measures are necessary to elucidate developmental processes leading to complex diseases.[39] But to date, small single-locus effect sizes have made it challenging to incorporate genetic information into ongoing cohort studies. To address the challenge of small effects, we used a multi-locus profile. The resulting genetic risk score enables measurement of a larger, genome-wide effect size and reduces the number of hypothesis tests to one, making follow-up of GWAS findings tractable in cohort studies that are needed to study development. Third, the longitudinal results illustrate that investigations of obesity as an outcome to developmental processes can inform public health initiatives and research priorities by identifying specific phases in development when genetic risk becomes manifest and thus might be amenable to intervention. Childhood growth in general, and in particular growth during the period between birth and the adiposity rebound, should be a focus for future research to understand genetic contributions to the development of obesity.

We acknowledge three limitations. First, we derived our genetic risk score from GWAS of Europeans and conducted our study in European-descent individuals; these results may not generalize to other populations.[5] Second, our family histories included only parents. It is possible more complete family histories have greater overlap with the genetic risk score. Third, we were unable to characterize growth trajectories during the earliest stages of life; regular follow-up of the cohort did not begin until age 3 years. However, results from our analyses of birthweight and of birth-3 weight gain were consistent with previous genetic investigations of this interval that did include repeated measurements.[109, 110, 153, 154] Moreover, we were able to capture growth from age 3 years onwards with a high degree of resolution; our study included 12 measurements taken over the subsequent 35 years. In addition to repeated measures of height and weight, our study included more direct measures of adiposity, including childhood measurements of skinfold thicknesses and adult measurements of waist circumference and fat mass, all of which were associated with our genetic risk score in parallel to BMI. Thus, the results present compelling evidence that SNPs identified in GWAS of adult BMI and other obesity-related phenotypes predispose to more rapid growth in childhood, leading to increased risk for obesity in adulthood, and provide information not forthcoming from a simple analysis of family history.

Table 3.1. Descriptive Statistics and Correlations with Genetic Risk Score and Family History Score for Anthropometric Assessments Among Individuals of European Descent (n=856). All correlations were adjusted for sex. Tests of statistical significance were conducted using heteroskedasticity robust standard errors. *** p<0.001, **p<0.01, *p<0.05.

Measure / Age at Measurement	Mean	SD	Correlation (r) with Genetic Risk Score	Correlation (r) with Family History Score
Body Mass Index (kg/m^2)				
3	16.33	(1.28)	0.08 *	0.11 **
5	15.88	(1.18)	0.13 ***	0.19 ***
7	15.82	(1.29)	0.17 ***	0.24 ***
9	16.33	(1.62)	0.18 ***	0.21 ***
11	17.49	(2.09)	0.16 ***	0.22 ***
13	19.59	(2.47)	0.18 ***	0.23 ***
15	20.37	(2.60)	0.17 ***	0.30 ***
18	22.73	(3.02)	0.15 ***	0.31 ***
21	23.74	(3.41)	0.18 ***	0.31 ***
26	24.89	(4.27)	0.16 ***	0.29 ***
32	26.03	(4.80)	0.13 ***	0.31 ***
38	26.92	(5.13)	0.14 ***	0.34 ***
Alternative Measures of Adiposity				
Subcapular Skinfold Thickness (mm)				
7	5.87	(1.92)	0.07 **	0.12 **
9	6.67	(2.80)	0.11 ***	0.10 **
Tricep Skinfold Thickness (mm)				
7	8.46	(2.46)	0.07 *	0.10 **
9	11.06	(4.12)	0.09 **	0.10 **
Waist Circumference (mm)				
26	798.68	(96.40)	0.14 ***	0.24 ***
32	841.81	(110.95)	0.11 ***	0.25 ***
38	858.39	(122.84)	0.12 ***	0.26 ***
Fat Mass (kg)				
32	21.41	(10.54)	0.10 **	0.23 ***
38	23.59	(10.96)	0.11 **	0.27 ***
Gestational and Childhood Growth				
Birthweight (kg)	3.38	(0.52)	0.00	0.07 *
Weight Gain Birth - 3 years (kg)	11.31	(1.52)	0.09 *	0.02
Adiposity Rebound				
Age (years)	6.11	(1.10)	-0.13 ***	-0.21 ***
Body Mass Index (kg/m^2)	15.57	(1.00)	0.17 ***	0.23 ***

Table 3.2. Polygenic Risk for Adult Obesity is Mediated by Developmental Phenotypes of Rapid Early Growth. Panel A presents the analysis of mediation of genetic risk for obesity by birth-3 weight gain. Panel B presents the analysis of mediation of genetic risk for obesity by age and BMI at adiposity rebound. The columns labeled "Bivariate Models" present bivariate effects (relative risks and 95% confidence intervals) for the genetic risk score and the developmental phenotypes from Poisson regression models. The columns labeled "Multivariate Model" present the independent effects of the genetic risk score and the developmental phenotypes from multivariate Poisson regression models. Mediation analyses are reported in the final row of each panel. Mediation Ratios were calculated from the indirect and direct effects estimated from structural equations (**Appendix B, Supplemental Table 3**). The mediation ratio describes how much of the effect of genetic risk is mediated by the developmental phenotype. All analyses were adjusted for sex and included the n=856 European-descent individuals in the analysis sample. Birth-3 weight gain and adiposity rebound measures were standardized to have means of 0 and standard deviations of 1 for analyses.

	Age Range / % Ever Obese (n)							
	Obesity in the Teenage Years 5% (n=47)		**Obesity in the Twenties** 11% (n=96)		**Obesity in the Thirties** 22% (n=191)		**Chronic Obesity (ages 15-38 years)** 8% (n=72)	
Panel A. Polygenic Risk for Obesity is Partly Mediated by Weight Gain Between Birth and Age 3 Years								
	Bivariate Models	Multivariate Model	Bivariate Models	Multivariate Model	Bivariate Models	Multivariate Model	Bivariate Models	Multivariate Model
Genetic Risk Score	1.39 (1.08, 1.81)	1.30 (1.00, 1.69)	1.37 (1.13, 1.68)	1.34 (1.10, 1.63)	1.23 (1.08, 1.39)	1.21 (1.07, 1.37)	1.37 (1.09, 1.74)	1.32 (1.05, 1.67)
Birth-3 Weight Gain	1.78 (1.38, 2.30)	1.72 (1.33, 2.22)	1.32 (1.10, 1.59)	1.28 (1.06, 1.54)	1.15 (1.02, 1.31)	1.13 (1.00, 1.28)	1.44 (1.15, 1.80)	1.39 (1.12, 1.73)
Mediation Ratio Sobel Test of Mediation		0.16 **p=0.021**		0.07 p=0.057		0.06 p=0.118		0.10 **p=0.037**
Panel B. Polygenic Risk for Obesity is Partly Mediated by Age and Body Mass Index at Adiposity Rebound								
	Bivariate Models	Multivariate Model	Bivariate Models	Multivariate Model	Bivariate Models	Multivariate Model	Bivariate Models	Multivariate Model
Genetic Risk Score	1.39 (1.08, 1.81)	1.18 (0.93, 1.49)	1.37 (1.13, 1.68)	1.20 (1.00, 1.45)	1.23 (1.08, 1.39)	1.14 (1.01, 1.29)	1.37 (1.09, 1.74)	1.20 (0.96, 1.50)
Age (years) at Adiposity Rebound	0.57 (0.48, 0.68)	0.57 (0.48, 0.68)	0.66 (0.58, 0.75)	0.66 (0.58, 0.75)	0.77 (0.70, 0.85)	0.78 (0.70, 0.85)	0.66 (0.57, 0.76)	0.66 (0.57, 0.76)
Body Mass Index at Adiposity Rebound	2.13 (1.70, 2.66)	2.09 (1.66, 2.64)	1.61 (1.36, 1.89)	1.56 (1.32, 1.85)	1.35 (1.20, 1.51)	1.33 (1.19, 1.49)	1.72 (1.44, 2.05)	1.68 (1.39, 2.02)
Mediation Ratio Sobel Test of Mediation		0.58 **p<0.001**		0.44 **p<0.001**		0.41 **p<0.001**		0.45 **p<0.001**

Figure 3.1. Developmental phenotypes of rapid early growth hypothesized to mediate polygenic risk for obesity. The genetic epidemiology of obesity indicates that a large number of common polymorphisms each contribute small, additive increments to risk for obesity.[13, 89] The combined influence of these polymorphisms can be summarized in a polygenic risk profile.[65] The developmental epidemiology of obesity highlights three developmental phenotypes of rapid early growth that predispose children to become obese in later life: (1) growth during gestation, (2) postnatal growth, and (3) adiposity rebound.[134] [10] We tested the hypothesis that these developmental phenotypes would mediate polygenic risk for adult obesity.

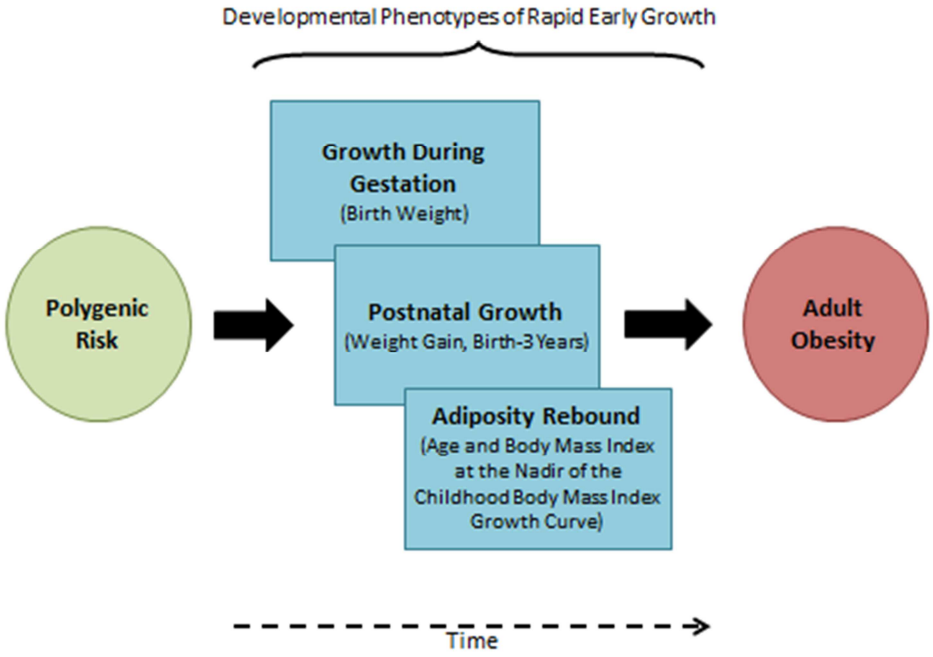


Figure 3.2. Individuals with higher obesity genetic risk scores (GRSs) were larger and grew more rapidly as children and as adults. The solid line represents the population mean trajectory (average genetic risk). Dashed lines are for subgroups +/- 1 standard deviation of the GRS (high and low genetic risk). Trajectories were derived from the life-course growth model (intercept fitted at age 13 years; linear and quadratic slopes fitted over ages 3-13 years and 13-38 years) including intercept and linear slope effects for the genetic risk score. Analyses included n=856 European-descent individuals.

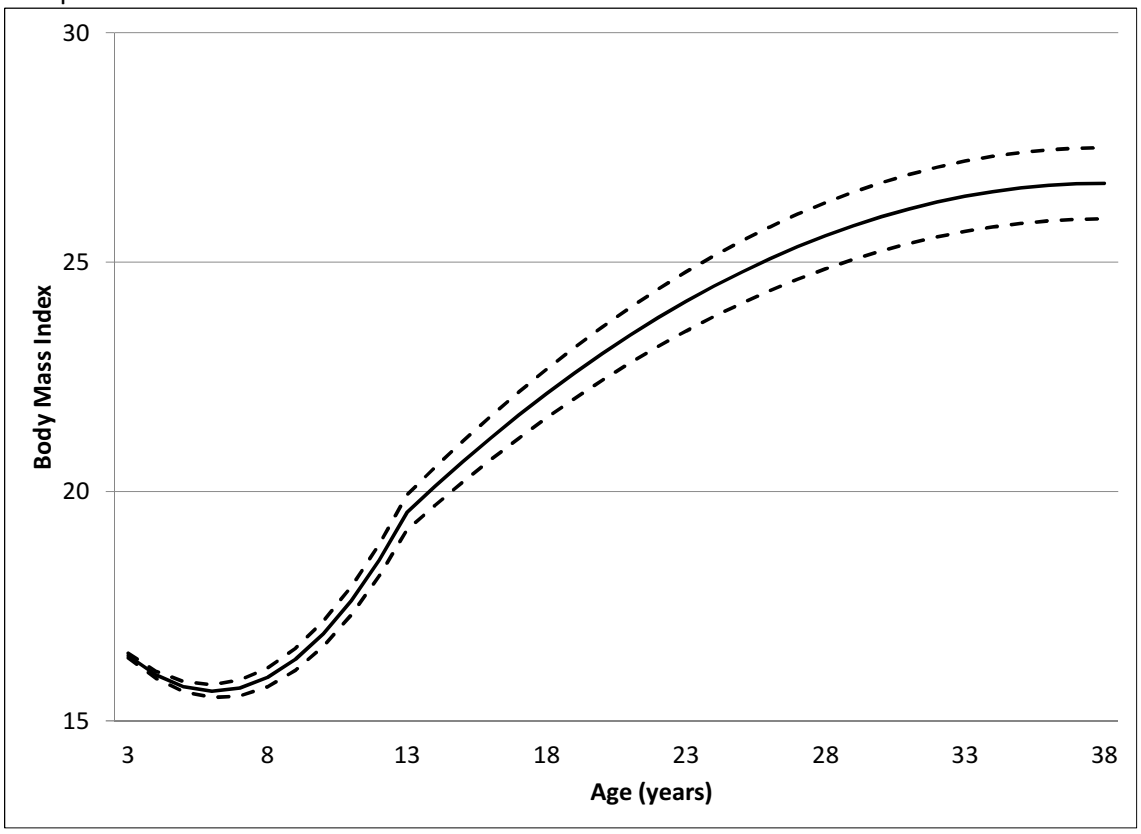


Figure 3.3. Individuals with higher genetic risk scores were more likely to be obese across two decades of adult follow-up. Error bars reflect 95% confidence intervals (CIs). The genetic risk score was dichotomized at the sample mean to create low and high genetic risk categories. Relative risks (RR) and 95% CIs are reported from Poisson regression models adjusted for sex including the n=856 European-descent individuals in the analysis sample.

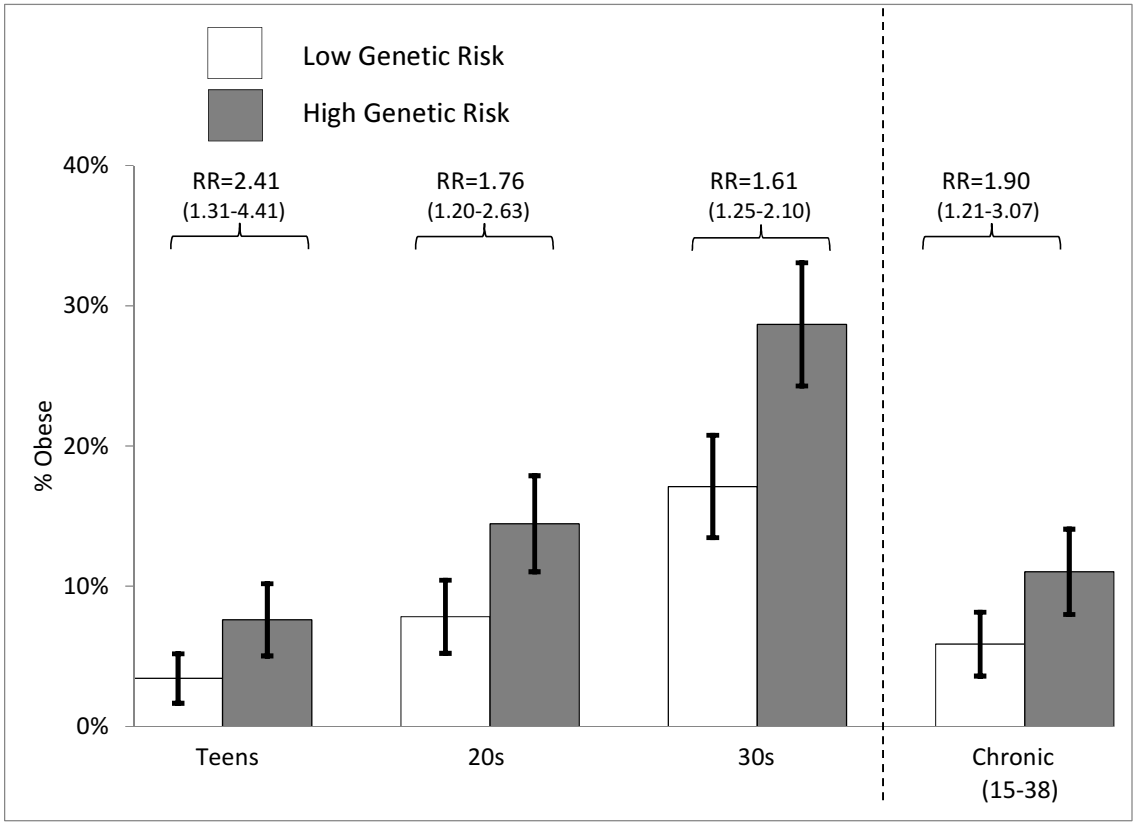
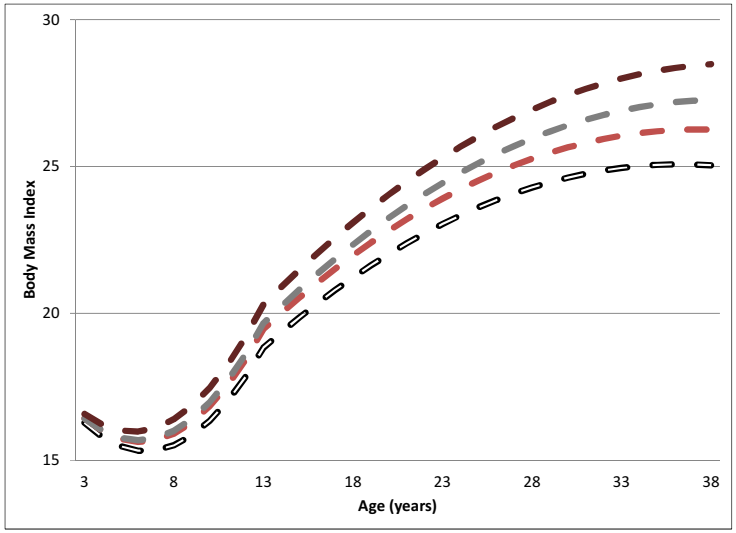
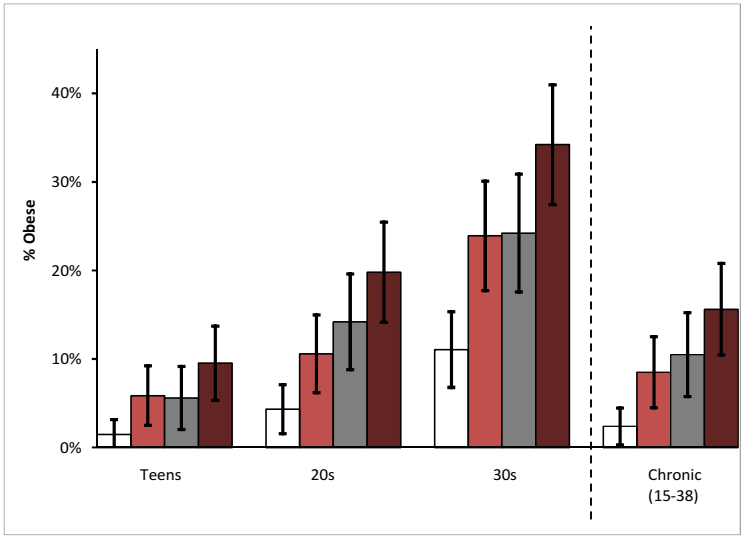


Figure 3.4. The genetic risk score contained information about children's growth and their obesity risk that was not available in their family histories. Genetic risk and family history made independent and additive contributions to life-course growth predictions and to adult obesity risk in n=856 European-descent individuals. **Panel A** shows that genetic risk and family history made additive contributions to growth predictions. **Panel B** shows that genetic risk and family history made additive contributions to children's risk of becoming obese. Error bars reflect 95% confidence intervals. Statistical analyses illustrating the independence of the genetic risk score and family history in predicting growth and obesity risk are presented in **Supplementary Table 2 (Appendix B)**.





Low Familial Risk, Low Genetic Risk

Low Familial Risk, High Genetic Risk

High Familial Risk, Low Genetic Risk

High Familial Risk, High Genetic Risk

CHAPTER 4.

POLYGENIC RISK ACCELERATES THE DEVELOPMENTAL PROGRESSION TO HEAVY, PERSISTENT SMOKING AND NICOTINE DEPENDENCE: EVIDENCE FROM A 4-DECADE LONGITUDINAL STUDY

INTRODUCTION

Cigarette smoking is a costly, prevalent public health problem. The US Centers for Disease Control and Prevention attribute 400,000+ deaths and $95 million in lost productivity to smoking during 2000-2004.[156] About 20% of adults still smoke daily despite widespread knowledge of smoking's health effects and increasing economic costs to smokers due to rising taxes.[157] Thus, more effective interventions to prevent smoking, motivate smoking cessation, and prevent relapse back to smoking are needed.[158-160]

Studies of twins suggest that genetic differences between individuals play an important role in smoking behavior, cessation, and in response to anti-smoking interventions.[161] Recent genome-wide association studies (GWAS) in adult smokers and former smokers revealed genes that relate with genome-wide significance to smoking quantity (number of cigarettes smoked per day).[162-164] These genes are already being used in clinical applications; e.g., to predict smoking cessation likelihood and in pharmacogenetic analyses.[165-169] An important additional step in the translation of these GWAS findings is to test if genetic markers that predicted smoking quantity in GWAS also predict the development of smoking behavior in adolescence.[170, 171] This question is of critical importance for public health practice as intervention to disrupt genetic risk is likely to be most effective early in the development of dependence. Important developmental phenotypes in the pathogenesis of adult dependence include smoking initiation, conversion to daily smoking during adolescence, and rapid progression to heavy smoking.[172] Early, rapid progression from smoking initiation to heavy use is a signal risk for adult nicotine dependence.[173-176] Therefore, this research tested relations of GWAS-identified genetic risk with both adolescent and adult smoking phenotypes and then determined the extent to which genetic effects on the former affected the adult phenotype outcomes.

In this study, we tested prospective associations between genetic risks and adolescent developmental and mature adult phenotypes of smoking behavior (**Figure 4.1)**. We examined genetic risks in the Dunedin Study, a birth cohort (n=1,037) followed to age 38 years with >90% retention. We collected smoking behavior data at 8 assessments spanning ages 11-38 years. This allowed us to study the effects of genetic risk in the cohort as members initiated smoking during adolescence, converted to daily smoking and progressed to heavy smoking during the teenage and young adult years, and as they developed nicotine dependence and struggled with cessation in their 20s and 30s. We tested whether individuals at higher genetic risk progressed more rapidly from smoking initiation to heavy smoking, if they smoked more heavily as adults, if they were more nicotine dependent, and if they were more likely to fail in their cessation attempts. Finally, we tested the hypothesis that genetic risk accelerates the developmental progression from smoking initiation to heavy smoking, and this, in turn, increases the severity of adult smoking problems such as heavy, intractable smoking and nicotine dependence. This model has relevance to public health interventions that might delay the developmental progression to heavy smoking.

METHODS

Sample

Participants are members of the Dunedin Multidisciplinary Health and Development Study, a longitudinal investigation of health and behavior in a complete birth cohort. Study members (N=1,037; 91% of eligible births; 52% male) were all individuals born between April 1972 and March 1973 in Dunedin, New Zealand, who were eligible for the longitudinal study based on residence in the province at age 3 and who participated in the first follow-up assessment at age 3. The cohort represents the full range of socioeconomic status in the general population of New Zealand's South Island and is primarily white. Assessments were carried out at birth and at ages 3, 5, 7, 9, 11, 13, 15, 18, 21, 26, 32, and, most recently, 38 years, when 1,007 Study members were still alive, with over 90% retention. At each assessment wave, study members are brought to the Dunedin research unit for a full day of interviews and examinations. The Otago Ethics Committee approved each phase of the study and informed consent was obtained from all study members.

Measures

Genetic Risk Score

A challenge for developmental research following-up GWAS discoveries is that effect sizes for individual single-nucleotide polymorphisms (SNPs) are small; the largest effects for smoking quantity approach a change of 1 cigarette per day per risk allele. Moreover, many of the longitudinal studies with data necessary to investigate developmental phenotypes are underpowered to test single-SNP effects.[54] However, there is evidence that smoking-associated loci make additive contributions to risk, recommending aggregating risk alleles.[59, 60, 177] Summing risk alleles across GWAS-identified SNPs to compute a "genetic risk score" (GRS) yields a quantitative index of genetic risk with a normal distribution[65] and a potentially larger effect size.

We derived the genetic risk score (GRS) from 3 recent meta-analyses of GWAS that used as their phenotype cigarettes smoked per day.[162-164] To construct the genetic risk score (GRS), we considered single-nucleotide polymorphisms (SNPs) from regions with genome-wide significant associations in at least two meta-analyses: All 3 meta-analyses identified SNPs in the q25.1 region of chromosome 15 containing the *CHRNA5-CHRNA3-CHRNB4* gene cluster. Two meta-analyses identified SNPs in the q13.2 region of chromosome 19 containing the gene *CYP2A6*. These genes influence nicotine response and nicotine metabolism, have been linked with nicotine dependence, and are candidate genes in research into the development of smoking behavior.[60, 178-185] Therefore, we focused our inquiry on GWAS-identified SNPs in these two regions. In 15q25.1, we selected the SNPs rs16969968, rs6495308, rs8032771, and rs12595538. The SNPs rs16969968 and rs6495308, which fall within the *CHRNA5-CHRNA3-CHRNB4* gene cluster, were shown previously to have independent associations with smoking quantity[163] (see also[186]). The SNPs rs8032771 and rs12595538, which are located downstream of the *CHRNA5-CHRNA3-CHRNB4* gene cluster, were in weak linkage-disequilibrium (LD) with rs16969968 and rs6495308 ($R^2 \leq 0.10$), and were genome-wide significant in the largest meta-analysis[162] ($p<1x10^{-16}$ for both; p-values for these SNPs were not published in the other two meta-analyses). In 19q13.2, we selected the SNPs rs7937 and rs4105144. We summed alleles associated with higher smoking quantity to calculate the GRS.

To validate this GRS, we used independent data from the Atherosclerosis Risk in the Communities (ARIC) sample, accessed through the NIH database of Genotypes and Phenotypes (dbGaP).[81] When a GRS SNP was not available in ARIC, we selected the closest LD proxy for that SNP to include in the GRS. Among European-descent ARIC participants, each standard deviation (SD) increase in

the GRS predicted a 0.99 pack-year increase in lifetime cigarette consumption among individuals who had ever smoked (p=0.001) and a 0.65 cigarette increase in daily consumption among these ever smokers (p=0.001).

Dunedin cohort genotyping was conducted with an Illumina BeadPlex Array using DNA extracted from whole blood (93% of the sample) or buccal swabs (7% of the sample). GRS SNPs or proxies (linkage $R^2 \geq 0.85$) were called successfully in >95% of European-descent study members (**Appendix C**, **Supplemental Table 4.1**). These n=880 individuals formed the analysis sample. Cohort members carried an average of 7.06 of 12 possible risk alleles (SD=2.27). The GRS was standardized to have mean=0 and standard deviation=1 for analyses (genetic risk Z-score).

Family History of Smoking

Family histories of smoking were available for 99% of the cohort. The family history consisted of reports of smoking history provided by study members and both parents for study members' siblings, parents, and grandparents. The family history was summarized as the proportion of family members in the pedigree who were ever regular smokers, adjusted to account for differences in genetic relatedness to the proband of first- and second-degree relatives.[187]

Smoking Behavior

The developmental progression of smoking behavior in the Dunedin cohort is described in **Figure 2 Panel A.** Measurement of adolescent developmental phenotypes and mature phenotypes of smoking behavior are described in **Figure 4.2 Panel B.**

Analyses

Data analysis was divided into three parts: First, we analyzed associations between the GRS and developmental phenotypes of smoking behavior. Second, we analyzed associations between the GRS and mature phenotypes. Third, we tested whether developmental phenotypes mediated associations between the GRS and mature phenotypes. We analyzed continuous data using ordinary least squares, count data using negative binomial, and categorical data using Poisson regression models. We used longitudinal data analysis techniques to account for differences in exposure time: We analyzed hazards of smoking initiation, progression to heavy smoking, becoming nicotine dependent, and relapsing from a quit attempt using Cox proportional hazard models. To account for differences in the frequency with which study members attempted cessation, we constructed panel datasets that included one observation per study member per assessment (for the age 18-32 data) and one observation per study

member per quit attempt (for the life-history calendar data). We used these panel datasets to analyze the genetic effect on smokers' risks of cessation failure during ages 18-32 years and on their hazards of relapse during ages 32-38 years. We accounted for non-independence of repeated observations of individuals using generalized estimating equation models of risks and conditional risk-set models of hazards.[188, 189] Analyses were adjusted for sex and conducted using Stata 11.0.[145] Panel-data models were fitted to longitudinal repeated-measures data using "XT" and "ST" commands in Stata 11.0. We evaluated mediation using structural equation modeling.[140, 142, 143] Unless otherwise noted, effect-sizes are presented for one-unit increases in the genetic-risk Z-score (GRS).

RESULTS

Genetic risk was not related to smoking initiation. The GRS was not associated with whether individuals initiated smoking or with the timing of initiation (relative risk (RR) for smoking initiation=0.98, 95% CI [0.95-1.02], cumulative hazard ratio (HR) for initiation=1.01, [0.94-1.09] based on a one-unit increase in GRS z-score). Subsequent analyses focused on the 627 Dunedin cohort members who initiated smoking at some point during follow-up (**Figure 4.2**).

Genetic risk was related to the progression of smoking behavior. Individuals at higher genetic risk were more likely to progress to smoking ≥20 cigarettes/day and did so more rapidly (HR=1.35 [1.14-1.58]). **Figure 3 Panel A** shows the cumulative hazards for smoking ≥20 cigarettes/day for individuals at low, average, and high genetic risk. An unexpected finding was that individuals who initiated smoking but who did not progress to daily smoking or to heavy smoking, so-called "chippers", were at the lowest genetic risk of any group in the cohort (**Figure 4.3 Panel B).**

Adolescents at higher genetic risk were more likely to convert to daily smoking early and to progress rapidly from initiation to smoking ≥20 cigarettes/day. Among ever-smokers, 19% converted to daily smoking by age 15 years (early conversion) and 10% progressed to smoking ≥20 cigarettes/day by age 18 years (rapid progression to heavy smoking). Each unit increase in the GRS predicted a 24% increase in the relative risk of early conversion (RR=1.24 [1.06-1.45]) and a 43% increase in the relative risk of rapid progression (RR=1.43 [1.10-1.86]).

Individuals at higher genetic risk smoked more heavily across the lifespan. Individuals at higher genetic risk accumulated more pack-years across 38 years of follow-up. Each one-unit increase in the GRS predicted an additional pack-year in lifetime cigarette consumption among ever-smokers (B=1.05

[0.36-1.73]) (**Figure 4.4 Panel A).** We also analyzed the persistence of heavy smoking as the number of assessments at which individuals smoked ≥20 cigarettes per day. Individuals at higher genetic risk smoked heavily at more assessments (incidence rate ratio (IRR) for number of assessments as a heavy smoker=1.26 [1.07-1.49]).

Smokers at higher genetic risk were more likely to develop nicotine dependence. Through age 38 years, 27% of ever-smokers developed nicotine dependence. Individuals at higher genetic risk were more likely to become nicotine dependent compared to individuals at lower genetic risk and were nicotine dependent at more assessments (HR for nicotine dependence =1.27 [1.09-1.47]; IRR for assessments with nicotine dependence=1.22 [1.06-1.41]) (**Figure 4.4 Panel B**).

Smokers at higher genetic risk were more reliant on smoking as a coping strategy. In addition to testing genetic associations with nicotine dependence, we also asked whether cohort members at higher genetic risk were more reliant on smoking to cope with stress. Among study members who smoked daily during ages 32-38 years (n=261), those at higher genetic risk relied more heavily on smoking as a coping strategy (B=0.22 [0.10-0.33]).

Smokers at higher genetic risk were more likely to experience cessation failure. Assessment of cessation failure is challenging.[190] Therefore, we looked for convergent evidence across two approaches to testing genetic associations with cessation failure. We first analyzed study members' reports of cessation failure between ages 18-32 years. Across 14 years of follow-up, n=405 cohort members smoked daily. 90% of this group made at least one quit attempt and 51% reported a cessation failure at one or more assessments. Cohort members at higher genetic risk were more likely to experience cessation failure in their quit attempts (RR=1.11 [1.01-1.22]).

We next used the month-to-month life history calendars to look closely at cohort members' smoking behavior during their 30s, when cessation was most common. Across 96 months of follow-up, n=261 cohort members smoked daily. 52% of these smokers made a quit attempt lasting one month or more. Relapse was common (occurring in 61% of quitters). Quitters at higher genetic risk were more likely to relapse and did so sooner after quitting (HR=1.22 [1.02-1.45]). Only 19% of daily smokers achieved successful cessation (abstinent for ≥1 year through age 38). Smokers at higher genetic risk were less likely to have achieved successful cessation at the end of follow-up (RR=0.73 [0.57-0.93]) (**Figure 4.4 Panel C**).

Early conversion to daily smoking and rapid progression to heavy smoking mediated genetic associations with adult smoking problems. We derived an index of adult smoking problems from a principal components analysis of 3 indicators: a) pack-years smoked by age 38 years; b) total number of Fagerstrom symptoms across assessments; and c) the number of assessments at which study members reported cessation failure. The adult smoking problems factor explained 78% of the variance in the 3 indicators (factor loading for heavy smoking=0.61; for nicotine dependence=0.60; for cessation failure=0.52). Individuals at higher genetic risk developed more smoking problems in adulthood (r=0.10, p=0.012). We next tested whether this association was accounted for by the more rapid developmental progression of smoking behavior among individuals at higher genetic risk. 81% of this association was accounted for by the two adolescent developmental phenotypes of early conversion to daily smoking and rapid progression to smoking≥20 cigarettes/day (**Appendix C**, **Supplemental Table 4.2**). Among individuals who did not exhibit these adolescent developmental phenotypes of rapid smoking progression, genetic risk was uncoupled from the development of smoking problems in adulthood (r=0.05, p=0.176).

The genetic risk score captured information that could not be ascertained from a family history of smoking behavior. The family history score and the GRS were uncorrelated (r=0.011). Both family history and the GRS predicted study members' smoking phenotypes (**Table 4.1**). When family history and the GRS were both standardized and included in regression models simultaneously, GRS coefficients were unchanged and remained statistically significant. Thus, the GRS contained different information about risk for developmental and mature phenotypes of smoking behavior compared to family history.

COMMENT

Etiological research on substance abuse highlights the importance of progression from initiation to heavy use during adolescence in the development of dependence in adulthood.[47, 48] In this study, we linked the developmental progression of smoking behavior to genetic risk. We derived a genetic risk score (GRS) from GWAS of smoking quantity. This GRS was not related to smoking initiation. However, individuals at higher genetic risk did progress more rapidly from smoking initiation to heavy smoking. In fact, daily smokers who did not progress to heavy use were at lower genetic risk than individuals who never smoked. Critically, high genetic risk led individuals to become persistent heavy smokers, develop nicotine dependence, and struggle with cessation failure only to the extent that they progressed rapidly from smoking initiation to heavy smoking during adolescence.

Previous research has related polymorphisms in the genes included in our genetic risk score to developmental phenotypes of smoking behavior[60, 177, 181-184] and to mature phenotypes of adult smoking problems.[178-180, 191, 192] To our knowledge, ours is the first study to track the relations of particular genetic risk variants with the development of smoking behavior from initiation through conversion to daily smoking and progression to heavy smoking, and on to the mature phenotypes of persistent of heavy smoking, nicotine dependence, and struggles with cessation through mid-life. Moreover, this extended follow-up allowed us to show, for the first time, that GWAS-identified variation in 15q25.1 and 19q13.2 influences adult smoking problems through a pathway mediated by adolescent progression from smoking initiation to heavy smoking. Our study is also the first to show that GWAS-identified SNPs provide information about smoking risks that cannot be ascertained from a family history, including information about risk for cessation failure.

These findings should be considered in light of three limitations. First, the Dunedin Study sample consisted of European-descent individuals, as did the samples analyzed in the GWAS used to develop the GRS. Replication in other populations is needed.[193] Second, our analyses of cessation were subject to censored data. The life history calendars ended at the age 38 follow-up and thus the data do not reflect relations with phenotypic events occurring after this age. Also, self-reports of temporally remote events could be inaccurate due to forgetting or other biases. Third, the four-decades of follow-up in the Dunedin Study coincided with major secular events such as bans against smoking in the workplace. Comparisons of cohorts born at different times might elucidate gene-policy interactions in smoking behavior and reflect the generalizability of the current findings.[194]

Despite these limitations, this study has implications for etiological research and public health. With respect to etiology, our study makes 3 contributions: First, Next Gen sequencing studies and other efforts to ascertain causal variants responsible for GWAS signals may maximize their discovery potential by focusing on samples of young people strategically selected to reflect important developmental transitions. Such work could use experimental designs to test hypotheses about mechanisms of genetic risk on post-initiation phenotypes. Second, we demonstrated that a genetic risk score based on the assumption of additive risks can be used to follow-up GWAS results in a birth cohort far smaller than the original discovery samples. Future etiological research can use genetic risk scores to apply GWAS results to longitudinal studies. Third, results are consistent with the hypothesis in pediatric medicine that some adolescents, after only experimental use, are prone to quickly become heavy users and dependent.[11]

Turning to public health, our research adds a genetic dimension to long-standing arguments that early prevention could be a critical strategy in reducing cigarette consumption.[42] Specifically, our findings and others'[181] suggest that initiatives that disrupt the developmental progression of smoking behavior, such as surtaxes and age restrictions on tobacco purchases, may ameliorate some genetic risks.[195] Moving beyond population-level prevention, we showed that information about smoking risk captured in a score composed of GWAS-identified variants was independent of information that could be derived from a family history of smoking behavior. This novel finding suggests that genetic information could be used to identify "high-risk" youngsters for targeted prevention.[11, 196] However, the associations we detected between the genetic risk score and smoking phenotypes were small in magnitude. Small effect sizes do not preclude public health relevance,[197] but they do caution against the use of genetic information to evaluate risk in individuals;[7] children that our study would classify at high genetic risk are not guaranteed to become addicted if they try smoking and, even more importantly, children we would classify at low genetic risk are not immune to addiction. The public health use of the current findings must be tempered with recognition that most "risk-associated" genetic variation does not determine poor health outcomes and, correspondingly, its absence does not guarantee protection.[103, 198]

Table 4.1. Effect sizes for genetic and family history associations with developmental and clinical phenotypes of smoking behavior. Effect sizes are for a one standard deviation increase in the predictor variable (the genetic risk score or the family history score). Effect size measures are Pearson correlations, 'r'; relative risks 'RR' from Poisson regression models; incident rate ratios 'IRR' from negative binomial regression models; hazard ratios 'HR' from Cox proportional hazard models; and beta coefficients 'B', from linear regression models. All models were adjusted for sex. Effect-size measures with a star '*' were estimated from longitudinal datasets including repeated observations of individuals over time. Effect sizes in gray text were not statistically significant at the α=0.05 level.

	Effect Size Measure	Genetic Risk Score	Family History Score
Correlation between the Genetic Risk Score and the Family History Score	r	0.011 p=0.758	
Developmental Phenotypes			
Smoking Initiation (among n=880 Individuals; n=627 Who Ever Initiated Smoking)			
Ever-Smoker Status	RR	0.98 [0.95-1.02]	1.12 [1.07-1.17]
Lifetime Hazard for Smoking Initiation	HR*	1.01 [0.94-1.09]	1.06 [0.98-1.15]
Progression from Initiation to Heavy Smoking (among n=627 Ever-Smokers)			
Early Conversion to Daily Smoking (by Age 15 Years)	RR	1.24 [1.06-1.45]	1.52 [1.27-1.83]
Rapid Progression to Smoking ≥20 Cigarettes/Day (by Age 18 Years)	RR	1.43 [1.10-1.86]	1.68 [1.26-2.24]
Lifetime Hazard for Smoking ≥20 Cigarettes/Day	HR*	1.35 [1.14-1.58]	1.47 [1.23-1.76]
Mature Phenotypes			
Heavy Smoking Persistence (among n=627 Ever Smokers)			
Lifetime Cigarette Consumption (Pack Years)	B	1.05 [0.36-1.73]	2.49 [1.80-3.19]
Count of Assessments Smoking ≥20 Cigarettes/Day	IRR	1.26 [1.07-1.49]	1.49 [1.24-1.80]
Nicotine Dependence (among n=627 Ever Smokers)			
Lifetime Hazard to Becoming Nicotine Dependent (≥4 Fagerstrom Symptoms)	HR*	1.27 [1.09-1.47]	1.53 [1.29-1.80]
Count of Assessments with Nicotine Dependence	IRR	1.22 [1.06-1.41]	1.50 [1.28-1.75]
Smoking to Cope with Stress (Ages 32-38 Years, Among n=261 Daily Smokers)			
Smoking to Cope Score	B	0.22 [0.11-0.32]	0.09 [-0.05-0.23]
Cessation Failure			
Ages 18-32 years (n=405 Daily Smokers; among n=364 who Attempted Cessation)			
Risk of Cessation Failure	RR*	1.11 [1.01-1.22]	1.11 [1.00-1.23]
Ages 32-38 Years (n=261 Daily Smokers; n=136 Who Quit for ≥1 Month)			
Hazard of Relapse Following Quit Attempts Lasting ≥1 Month	HR*	1.22 [1.02-1.45]	0.96 [0.79-1.17]
Likelihood of Successful Cessation (among daily smokers)	RR	0.73 [0.57-0.93]	0.94 [0.73-1.20]

Figure 4.1. Genetic risk and the developmental progression of smoking behavior. In the hypothesized model, genetic risk influences the mature phenotypes of heavy smoking persistence, nicotine dependence, and cessation failure through a pathway mediated by three developmental phenotypes: smoking initiation, conversion to daily smoking; and progression to heavy smoking.

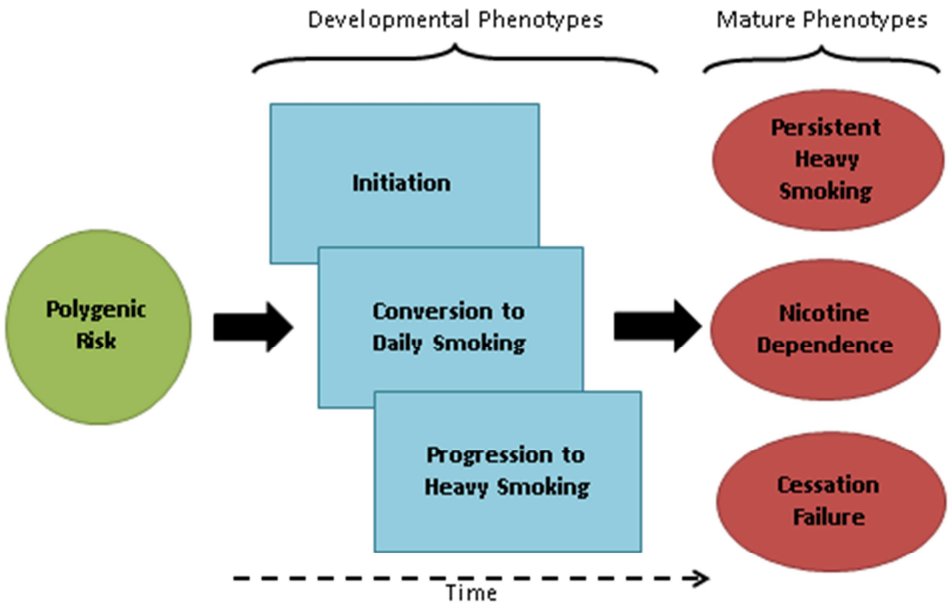
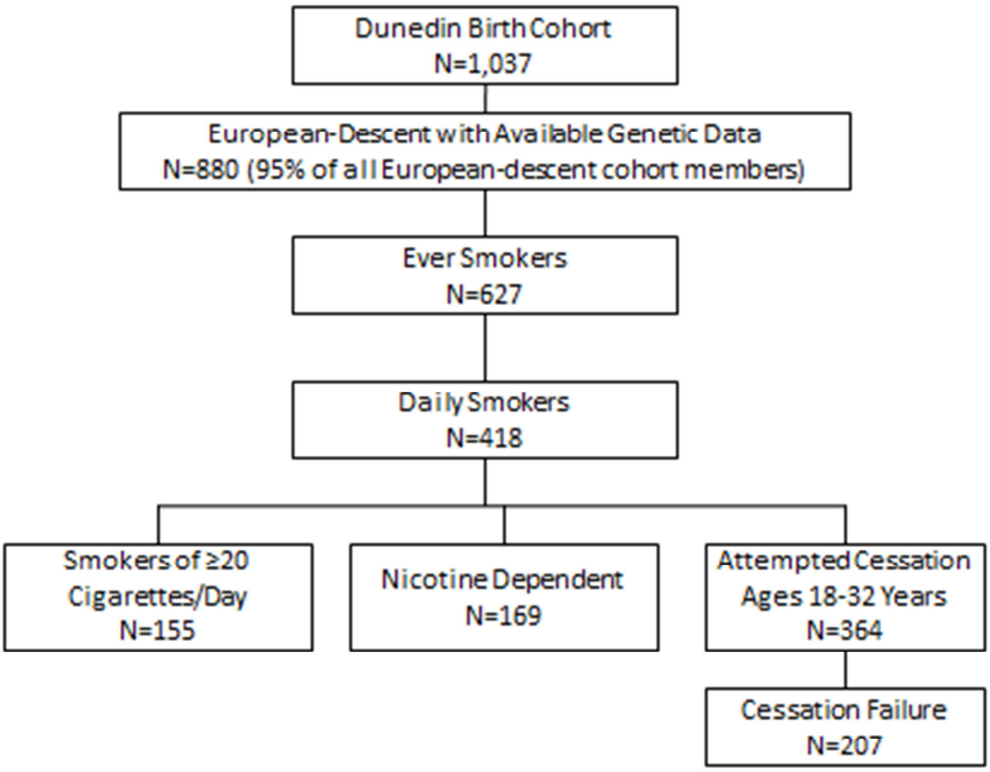


Figure 4.2. Smoking behavior in the Dunedin cohort.

Panel A. Developmental Progression of Smoking Behavior in the Dunedin cohort. Study members reported their smoking status during in-person assessments at ages 11 (percent ever-smokers=7%), 13 (13%), 15 (62%), 18 (66%), 21 (70%), 26 (70%), 32 (71%), and 38 years (71%) and their daily cigarette consumption at ages 13 (percent daily smokers=1%), 15 (14%), 18 (31%), 21 (34%), 26 (35%), 32 (30%), and 38 years (20%). We assessed nicotine dependence using the Fagerstrom Test of Nicotine Dependence (FTND),[199] completed by study members at the age-21, -26, and -38 assessments. We assessed cessation failure using study members' reports of quit attempts and outcomes at the ages 18, 21, 26, 32, and 38 assessments.



Panel B. Measurements of Developmental and Mature Smoking Phenotypes

Developmental Phenotypes	N	%
Initiation. Age at which study members reported first smoking at least occasionally. Survival time to initiation was calculated as the age at which a study member first began smoking at least occasionally. The 627 cohort members who initiated smoking represent 71% of the n=880 European-descent cohort members with available genetic data.	627	
Conversion Daily Smoking. The first assessment at which a study member smoked ≥1 cigarette per day.	418	67%
Early Conversion to Daily Smoking. Daily smoker at the age 15 assessment. Most study members who ever converted to daily smoking did so by the age-18 assessment (74%). Therefore, we defined "early conversion" to daily smoking as having converted by the previous assessment at age 15 years.	121	19%
Progression to Smoking ≥20 Cigarettes/Day. Ever a smoker of ≥20 cigarettes per day, ages 13-38 years. "Survival time" to heavy smoking was calculated as the number of years between initiation and the first assessment at which a study member smoked ≥20 cigarettes per day.	155	25%
Rapid Progression to Heavy Smoking. Smoker of ≥20 cigarettes per day by age 18 years. Most study members who ever became heavy smokers progressed to smoking ≥20 cigarettes/day by the age-21 assessment (63%). Therefore, we defined "rapid progression" as progression to smoking ≥20 cigarettes/day by the previous assessment at age 18 years.	61	10%
Mature Phenotypes	**M SD / N %**	
Lifetime Cigarette Consumption (Pack Years). Pack-years = the number of cigarettes smoked per day, divided by 20 and multiplied by the number of years smoked at that rate.[38] The mean and standard deviation of pack-years was calculated for ever-smokers.	8.39	8.95
Nicotine Dependence. Study members completed the Fagerstrom Test of Nicotine Dependence (FTND)[199] at the age-21, -26, and -38 assessments. The FTND was developed to measure the construct of physical dependence and includes the facets of needing to smoke in the morning to alleviate overnight withdrawal, needing to smoke many cigarettes per day, and invariance in smoking behavior, e.g. in the face of illness.[200] The FTND produces a "Fagerstrom score" score ranging from 0-10. Nicotine dependence is defined as a Fagerstrom score ≥4.[201, 202] We calculated survival time as years between smoking initiation and the first assessment at which a study member was nicotine dependent. Percent reflects ever-smokers who became nicotine dependent.	169	27%

Smoking to Cope with Stress. Study members were interviewed about how they coped with stress associated with their relationships, work, and finances at ages 32 and 38 years. Study members rated the extent to which they used different coping strategies. One of these strategies was smoking more. Ratings of smoking as a coping strategy were averaged and the average was standardized to produce a score with mean=0 and standard deviation=1.	0	1
Cessation Ages 18-32 years. To assess cessation failure prior to the period covered by the life history calendar, we used study member's reports collected at ages 18, 21, 26, and 32 years of whether they had made a quit attempt in the past year and whether they had failed in their quit attempt within one month. N reflects study members who attempted cessation at least once. Percent is calculated from study members who were ever daily smokers between ages 18 and 32 years (n=405).	364	90%
Cessation Ages 32-38 years. At the age-38 assessment, study members completed life history calendars[203] detailing their smoking behavior during each month from age 32 to age 38 years. Embedding recall of smoking behavior in a life-history calendar improves accuracy.[204] We used these data to identify 2 phenotypes: *Relapses* were ≥1 months of abstinence followed by ≥1 months of daily smoking; *Successful Smoking Cessation* was abstinence for ≥1 year through the time of the age-38 interview. N reflects study members who quit for ≥1 month between ages 32 and 38 years. Percent is calculated from study members who became daily smokers by age 32 and were daily smokers for ≥1 month during ages 32-38 years (n=261).	136	52%

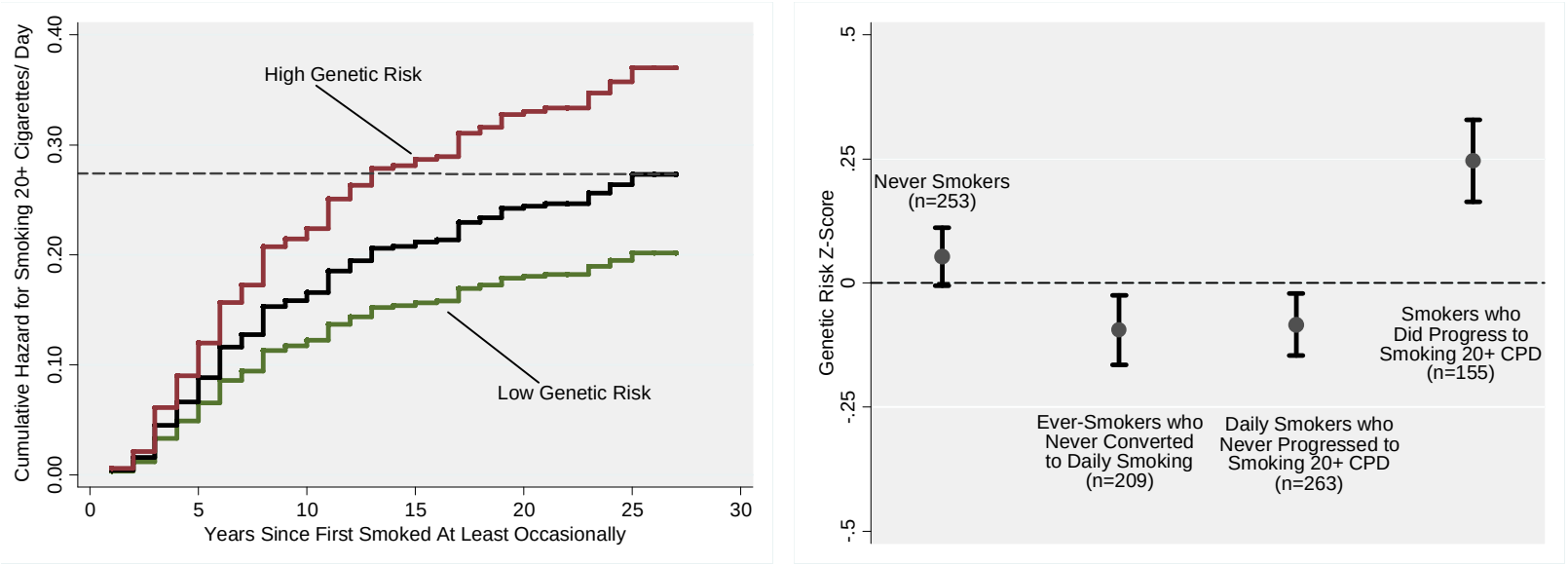
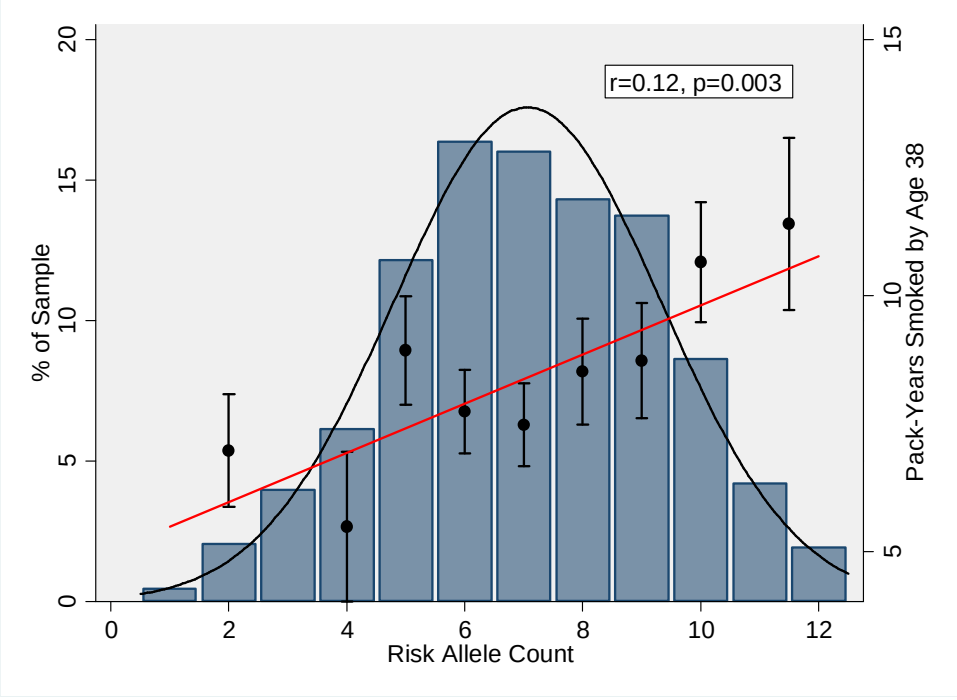


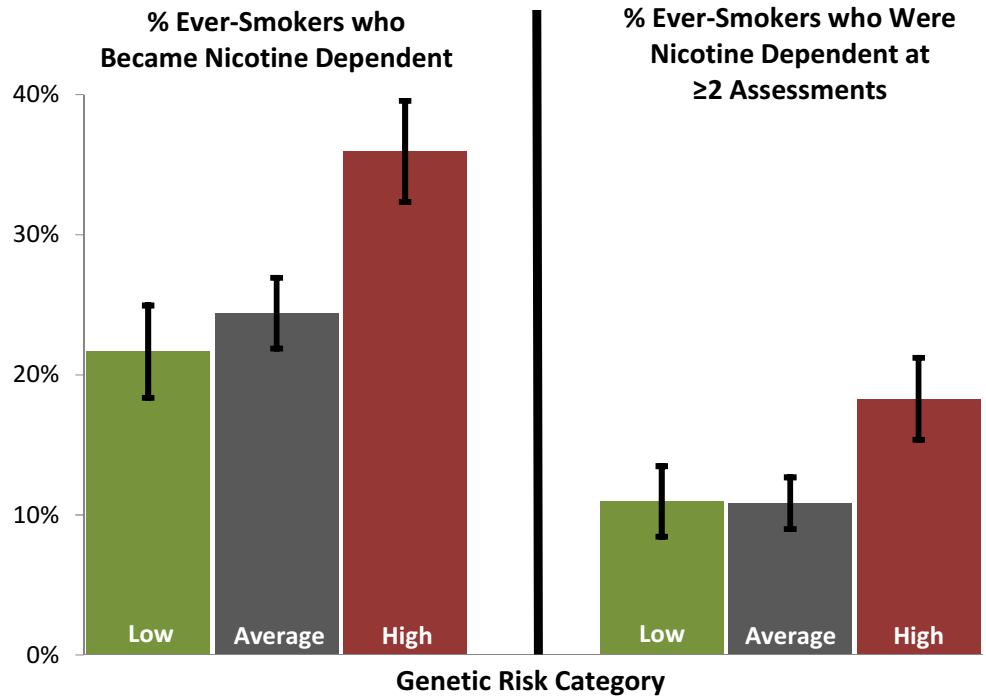
Figure 4.3. A genetic risk score derived from GWAS of smoking quantity is associated with the developmental progression of smoking behavior in a birth cohort of European-descent individuals. Panel A shows that individuals at higher genetic risk progressed more rapidly from smoking initiation to heavy smoking. Panel A graphs hazard functions for onset of heavy smoking among individuals at low genetic risk (genetic risk Z-score=-1, green line), average genetic risk (genetic risk Z-score=0, black line), and high genetic risk (genetic risk Z-score=1, red line). The dashed gray line marks the cumulative hazard for individuals at average genetic risk. The hazard function was estimated from a Cox proportional hazard model with time since onset of ever-smoking as the exposure time and the first assessment a study member reported smoking ≥20 cigarettes/day as the failure event. The hazard model included all individuals who ever initiated smoking (N=627). Individuals at higher genetic risk progressed more rapidly from smoking initiation to smoking ≥20 cigarettes/day (Hazard Ratio=1.35 [1.15-1.59]). Panel B shows that genetic risk was highest among individuals who progressed to heavy smoking and lowest among individuals who initiated smoking but who did not progress to heavy smoking. The figure shows the genetic risk Z-sores (+/- 1 standard error) for each group. "CPD" is "cigarettes per day." A genetic risk Z-score of 0 corresponds to the average genetic risk in the cohort. Error bars reflect standard errors of the sub-group means.

Figure 4.4. Genetic risk predicts mature phenotypes of smoking behavior. Panel A shows that among individuals who initiated smoking, those at higher genetic risk smoked more cigarettes by age 38 years. Ever-smokers were all individuals who initiated smoking by age 38 years (N=627). The bars of the histogram graph the percentages of the sample carrying 1-12 risk alleles. The dots and standard-error bars reflect average lifetime cigarette consumption (in pack-years) for ever-smokers carrying 1-3, 4, 5, 6, 7, 8, 9, 10, and 11-12 risk alleles. The regression line shows the association between the genetic risk score and pack-years smoked by age 38 years (Pearson Correlation r=0.12, p=0.003). Panel B shows that ever-smokers at higher genetic risk were more likely to be nicotine dependent. The bars of the chart graph the proportion of ever-smokers (n=627) at low, average, and high genetic risk (left side) who became nicotine dependent (≥4 Fagerstrom symptoms) by age 38 years; and (right side) who were nicotine dependent at two or more assessments. Panel C shows that smokers at higher genetic risk were more likely to experience cessation failure during their 30s. The bars of the chart graph the proportions of daily smokers at low, average, and high genetic risk (left side) who experienced relapse following a quit attempt lasting ≥1 month; and (right side) who achieved successful cessation (abstinence ≥1 year) through age 38 years. Percent with relapse was calculated from cohort members who quit smoking for ≥1 month during ages 32-38 years (n=136). Percent with successful cessation was calculated for cohort members who smoked daily during their 30s (n=261). In panels B and C, low genetic risk individuals had GRSs more than 1 standard deviation below the cohort mean; average genetic risk individuals had GRSs within 1 standard deviation of the cohort mean; and high genetic risk individuals had GRSs more than 1 standard deviation above the cohort mean. Error bars reflect standard errors.

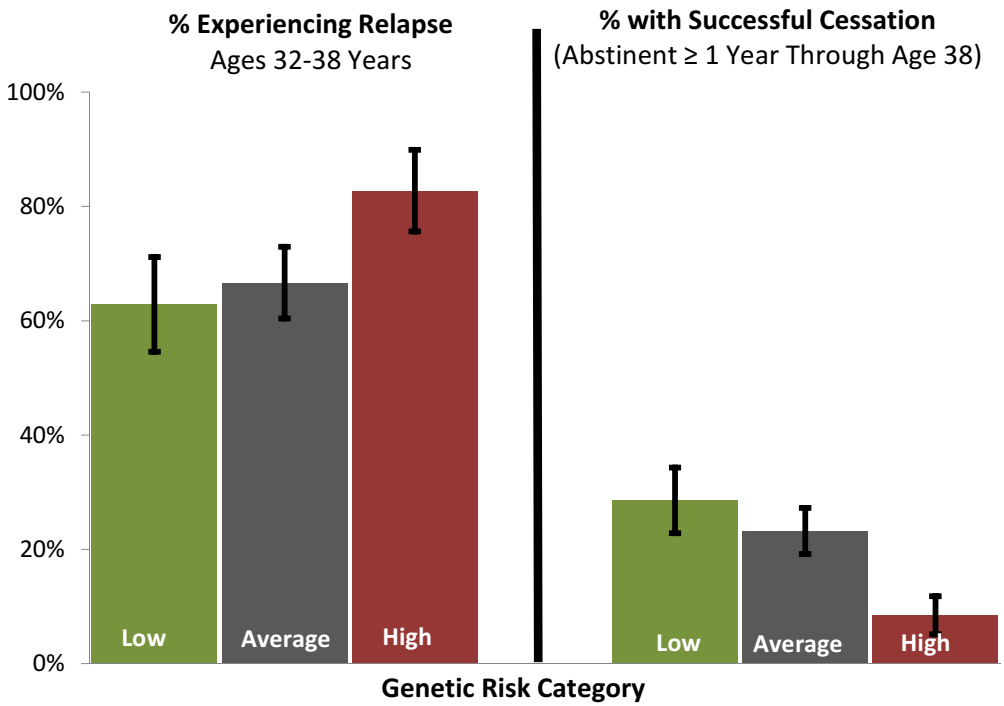
Panel A.



Panel B.



Panel C.



CHAPTER 5.

SUMMARY & CONCLUSIONS

The research described in the three preceding empirical chapters characterizes the public health implications of GWAS-discovered genetic risks for two prevalent and costly sources of morbidity and early mortality, obesity and smoking. I used theory-free genetic discovery research (genome-wide association studies, "GWAS") to derive multi-locus profiles of genetic risk. I next validated these "genetic risk scores" (GRSs) using data from a large, population-based cohort of older adults, the Atherosclerosis Risk in the Communities Study (n=15,792). I then used these GRSs to investigate how genetic risk influenced the development of obesity and smoking problems in a representative birth cohort (n=1,037) followed through their fourth decade of life. Results revealed that (1) GRSs can be used to investigate GWAS-discovered genetic risks for obesity and smoking in population-based cohorts much smaller than the original GWAS discovery samples; (2) genetic risks identified in GWAS manifest early in the development of obesity and smoking through processes that may be amenable to public health intervention; (3) the magnitudes of risk that can be predicted using genetic information are small; but (4) the risk information provided by genetic markers is independent of information available in a family history. These findings have three primary implications: First, they recommend of caution in the application of genetic information to predict health risks in individuals,[7-9] but suggest promise as more powerful but less common genetic risks are discovered in the continuing evolution of genomic research. Second, findings suggest that developing samples of children and adolescents should be a focus in genetic discovery research. Third, findings and add a genetic dimension to arguments for early intervention and may inform the best strategies likely to prevent obesity and smoking.[10, 11]

I. Implications for Research and Public Health

A. Genetic risk scores composed of genome-wide association study discoveries can provide valid measures of inherited risk for obesity and smoking. GRSs for obesity and smoking derived by summing risk-associated alleles across panels of GWAS-discovered SNPs are associated with quantitative traits (e.g. body mass index, smoking quantity) and clinical endpoints (e.g. obesity, nicotine dependence). These associations can be reliably detected in samples far smaller than those used for discovery research. In Chapter 3, for example, associations between the smoking GRS and cessation outcomes were detected in samples of only a few hundred smokers making quit attempts. Thus, GRSs provide a means to follow-up GWAS results in samples that are far smaller than those used for discovery research.

The validity of a genetic risk score may be specific to the population investigated in the GWAS used to construct the score. GWAS rely on "linkage", the correlation of spatially proximate nucleotides in the genome, to leverage measurements of one or two million SNPs to infer variation in the entire 15 million plus SNPs in the human genome.[205] The theory behind GWAS is that linkage between measured SNPs and unmeasured SNPs is strong enough that the "signal" from an un-measured SNP that is a causal factor in disease etiology can be detected through a "proxy" SNP that is measured.[206] Thus, GWAS-identified SNPs may be proxies in linkage with causal variants.[70] Patterns of linkage vary across population bottlenecks in human evolution, e.g. the migration out of Africa.[207] Therefore, GWAS results may not be consistent across ethnic groups defined by these population bottlenecks.[208] In Chapter 1, the obesity GRS was a robust predictor of body mass index and obesity in European-descent cohort members, but not in African-descent cohort members. Public health research using GRSs must be careful to match the population used for discovery with the population of interest or must validate the GRS in a new population before undertaking further etiological research.

B. The magnitudes of risks measured by GWAS-derived genetic risk scores for obesity and smoking are small at the level of the individual, but have implications for public health. Many ARIC and Dunedin cohort members at low genetic risk became obese or developed smoking problems. Similarly, many members of these cohorts at high genetic risk remained thin or never developed smoking problems. In the population based cohorts examined in Chapters 1-3, the

GRSs explained roughly 1% of population variance in the relevant outcomes. Consistent with prior research,[209] analyses in Chapter 1 show that a GRS explaining such a small fraction of variance in a health trait or behavior is not useful for individual prediction. Yet, the risks predicted by the GRSs were far from trivial. In analyses presented in Chapter 2 each 1-unit increase in the obesity GRS predicted a 40% increase in risk for chronic adult obesity. In analyses presented in Chapter 3, each 1-unit increase in the smoking GRS predicted a 30% increase in the hazards for becoming a heavy smoker and for developing nicotine dependence. Thus, these GRSs predict risks that are comparable to or in excess of risks that can be predicted by many other biomarkers considered of interest in the context of public health research.[210] Interventions that effectively ameliorated substantive portions of the genetic risks measured by these GRSs could reduce population rates of obesity and smoking and generate substantial improvements in health and substantial reductions in healthcare spending.[197]

C. Information captured in genetic risk scores for obesity and smoking is different from information that can be obtained from a family history. To have public health utility, measured genetic variation must provide new information over and above that already available in family history.[67] Therefore, this dissertation research tested the independence of risk scores derived from GWAS discoveries and from family histories. Chapters 2 and 3 report tests of the independence of the GRSs from risk scores derived from the body mass indexes of children's parents (for obesity) and the smoking behavior of their parents, siblings, and grandparents (for smoking). Correlations between GRSs and family-history scores were near zero and were not statistically significant. Regression analyses showed that the smoking and obesity GRSs provided information about risk that was independent of and additive to information that could be obtained from a family history. Thus, in the case of the smoking and obesity phenotypes, the genetic variants discovered in GWAS provide new information about risk over and above family history.

Risk effects estimated for family history were consistently larger than risk effects estimated for the GRSs, but the differences were small. Therefore, as new genetic discoveries are made, it may be useful to revisit the application of genetic information in the context of prospective risk assessment, particularly as a means to augment family history-based risk assessments.

A further note on the unexpected finding that genetic risks discovered in GWAS are un-related to family histories of obesity and smoking. The GRSs used in this dissertation research,

like family-history scores, explain only small fractions of the estimated genetic variance in the traits studied (1-4%). Family studies suggest that 50% or more of population variation in body mass index and smoking may be attributed to genetic factors.[13, 14] Therefore, the lack of overlap in the few percent of variance explained by each of the genetic risk and family history scores is not impossible. Nevertheless, it is somewhat unexpected that GWAS-discoveries should be completely unrelated to family histories.

There are three reasons that GWAS-derived GRSs and family-history scores might show so little overlap and these reasons relate to ongoing debate about the genetic architecture of common health problems. The debate concerns the importance of common genetic variation (like the SNPs measured in GWAS) as compared to rare mutations (variants that occur in <1% of the population) in causing common health problems.[35] GRSs derived from GWAS capture risk arising only from common variation. (There is some debate about the role of rare variants generating synthetic associations in GWAS, but the balance of evidence suggests that most GWAS hits are not artifacts of rare variants.[211-213]) In contrast, family histories comprise all genetic variation, both rare and common, in addition to environmental factors shared by family members. Therefore, one reason that the GRSs and family history scores might not overlap is that the GRSs reflect common genetic variants and family-history scores reflect primarily rare genetic variants and environmental risks shared by family members. A second reason that the GRSs and family-history scores might not overlap is that family history scores reflect so-called "epistatic" interactions among large numbers of common and/or rare variants and interactions between genetic variation and environments shared by family members.[214] A third reason that the GRSs and family-history scores might not overlap is that family history scores do reflect common variation, but common variation with effects too small to be detected in GWAS. The "infinitesimal model"[35] of trait heritability posits that very large numbers of common genetic variants, each with very small effects, contribute to common diseases.[215] Under this infinitesimal model, many causal common variants will escape detection in GWAS unless discovery samples grow into the millions and beyond. To the extent that family history scores reflect the combined influence of many thousands of common variants with infinitesimal effect sizes, they may not overlap with GRSs composed of a handful of common variants with larger effect sizes. Which of these explanations is correct has implications for public health. Family history is established as a clinically valid measure of genetic risk.[216-218] If common SNPs contain entirely different information from family history, they may be particularly valuable in the context of risk

assessment. However, methods that are better able to address the influence of large numbers of SNPs with "infinitesimal" effects, e.g. whole-genome scores,[68] are needed to resolve this question.

One way researchers have tried to address the problem of measuring genetic risks arising from individual variants with "infinitesimal" effects is to derive whole genetic risk prediction scores.[68, 219-221] Whole genetic risk prediction scores (whole genome scores) are derived by using large "training" datasets conduct GWAS and assign each SNP in a panel of 500,000-1 million a weight reflecting its individual contribution to a given health condition. The whole genome score is then calculated by generating a prediction based on the full panel of SNPs or an especially promising subset, e.g. the top 1,000 SNPs.[222, 223] These scores explain more of the variance common traits than GRSs composed of genome-wide significant SNPs.[58] E.g., for body mass index, a whole genome score based on 500,000 SNPs may explain as much as 15% of the population variance. As compared to the obesity GRS used in this dissertation, the whole genome score leverages a 15,000 fold increase in the number of genetic markers used to achieve between a 4- and 15-fold increase in predictive power. Thus, until genotyping costs decline further, this approach is unlikely to be cost-effective. Moreover, whole genome scores perform well in the samples from which they are derived, but much less well in independent replication samples.[223, 224] Therefore, their translational potential remains uncertain. Nevertheless, the whole genome approach is a promising direction for genetic risk prediction research. Investigating overlap in the risks predicted by whole genome scores and by family histories may provide preliminary guidance as to the value of whole-genome scores for clinical risk assessment.

D. Genetic risks for obesity and smoking discovered in genome-wide association studies manifest early in development. A critical question for understanding how genetic risk contributes to the etiology of obesity and smoking is to identify when in development genetic risk manifests.[171] This question is also central to translational science as intervention to ameliorate genetic risk may be most effective if deployed before health problems develop.[31] Research presented in chapters 2 and 3 shows that genetic risks identified in GWAS of adult samples first manifest earlier in life. Developmental phenotypes of obesity and smoking function as critical mediators of genetic risk. In the case of obesity, genetic risk manifests following birth and acts to influence childhood growth rates. The adiposity rebound, which is defined as the

nadir of the childhood body mass index growth curve, mediates roughly half the genetic effect on adult obesity. Children at higher genetic risk reached this nadir of adiposity earlier in life and at higher body mass index, and were subsequently more likely to become obese. Independent of their age and body mass index at adiposity rebound, children at high genetic risk were no more likely to become obese than their low genetic risk peers. In the case of smoking, genetic risk manifests following smoking initiation and acts to accelerate the developmental progression from experimentation to regular and heavy use during adolescence. Rapid developmental progression, defined as conversion to daily smoking by age 15 years or progression to smoking≥20 cigarettes/day by age 18 years, mediates the majority of the genetic effect on heavy smoking, nicotine dependence, and cessation failure. Individuals at higher genetic risk were more likely to progress rapidly in their smoking behavior and, in turn to become persistent heavy smokers, nicotine dependent, and unable to quit. Individuals at high genetic risk who did not progress rapidly from initiation to regular and heavy use during adolescence were no more likely to develop smoking problems in adulthood than their low genetic risk peers.

Results in Chapters 2 and 3 point to childhood and adolescence as critical stages in the manifestation of genetic risk and suggest new phenotypic targets for genetic research as well as highlighting possible windows that might be effectively targeted by public health interventions. Samples of children and adolescents, particularly those that include repeated measures of health outcomes and behaviors collected over time, will be critical to advancing understanding of the genetics of smoking and obesity.

Mediation analyses reported in Chapters 2 and 3 also highlight the importance of early intervention. Results presented in Chapter 2 suggest that genetic risks for obesity depend in part on rapid childhood growth. In parallel, results presented in Chapter 3 suggest that genetic risks for smoking problems depend on a rapid developmental progression from initiation to heavy use during adolescence. For individuals who did not exhibit these developmental phenotypes, genetic risks were uncoupled from adult health problems. Therefore, public health initiatives that encourage more healthy trajectories of growth in early childhood and that prevent or reduce smoking among adolescents may help to ameliorate some genetic risks.

II. Directions for Future Research

The research described in the three empirical chapters of this dissertation provides a foundation for a program of research to further the translation of genetic discoveries related to obesity and smoking risks to improve public health. This program of research includes three independent but related research objectives. The first objective is to understand the role of genetic risks in the etiology of socioeconomic disparities in health. Social gradients in health area major challenge for the healthcare system.[225, 226] Obesity and smoking are key mediators of these health disparities.[227] Therefore, to the extent that genetic discoveries can inform interventions to ameliorate social gradients in obesity and smoking, this would be an area of high translational impact. The second objective is to understand mechanisms in brain and in behavior that link genetic risk with health outcomes. Identifying mechanisms is a critical step in developing effective treatments, and therefore critical to the translation of genetic discoveries.[50, 77] The third objective is to extend the developmental research begun in this dissertation into the second half of the life course. Obesity and smoking are major contributors to morbidity and premature mortality in older adults.[228] This dissertation research establishes that GWAS discovered genetic risks influence how obesity and smoking develop in early life. The next step is to understand how genetic risks relate to changes in body mass index and smoking behavior in the second half of the life course.

Objective 1. To understand the role of genetic risks in the etiology of socioeconomic disparities in obesity and smoking. In developed countries, better educated and wealthier populations are less likely to become obese or to develop nicotine dependence relative to poorer and more poorly educated populations.[229, 230] This social gradient arises partly from a higher burden of environmental risks in lower socioeconomic strata, e.g. less access to health behavior promoting resources, targeted advertising of unhealthy foods and cigarettes. But a higher burden of discrete environmental risks does not fully account for observed social gradients in health.[231] Therefore, genetics constitute an important and as of yet unstudied potential source of these gradients. There are three possible roles that genetics could play in contributing to social gradients in health.[232] First, it is possible that genetic risks are more prevalent in lower socioeconomic strata. This is termed “gene-environment correlation” (rGE). Second, genetic risks may interact with environmental risks in a synergistic fashion. This is

termed “gene-environment interaction” (GxE). Third, genetic risks could be unrelated to social gradients in health.

Preliminary analyses of the Dunedin and ARIC cohorts confirm prior reports of strong social gradients in obesity and smoking.[100, 230] These analyses also reveal that genetic risks for obesity and smoking are similarly distributed across socioeconomic strata, i.e. there is no rGE between genetic risk and social class. Therefore, to the extent that genetic risks identified in GWAS contribute to socioeconomic disparities in obesity and smoking, they must do so in interaction with environmental exposures that do show a social gradient.

Recently, transdisciplinary research at the intersection of medicine, neuroscience, endocrinology, and immunology has begun to elucidate a model that describes how environmental stressors become biologically “embedded” in ways that contribute to the pathogenesis of chronic behavioral and physical health problems.[233] Such biological embedding of stress is likely to be critical to the etiology of health disparities: Disadvantaged populations face a higher burden of both severe acute stressors, e.g. violent victimization, and chronic stressors, e.g. food insecurity.[234] Thus, cumulative stress exposures may contribute indirectly to socioeconomic disparities in obesity and nicotine dependence through dysregulation of processes in body and brain that in turn influence patterns of activity, diet, and smoking behavior.[235] The gene-environment interaction hypothesis augments this biological embedding model by positing that certain individuals are more susceptible to the dysregulating effects of stress due to differences in their genes.[236]

One aim of my postdoctoral research will be to identify environmental stress exposures that may interact with polygenic risk in the etiology of health disparities. Criteria for candidate environmental moderators of genetic risk will (1) exhibit a social gradient; and (2) exert effects that coincide with the developmental stages identified in this dissertation as critical in linking genetic risks with adult obesity and smoking problems. In addition to these two scientific criteria, we will impose a third criterion: that the environmental exposure be amenable to public health intervention. The goal of this research is to devise means by which to disrupt the etiology of health disparities. Therefore, focusing on environments that can be modified through intervention is important to maximizing the translational potential of the research.

A second aim will be to test whether the stress exposures identified in the earlier aim alter the ways in which genetic risk manifests, either in terms of timing or in terms of magnitude. These analyses will utilize data from the Dunedin Study as well as from a new data source: the Environmental Risk in Development (E-Risk) Longitudinal Study, a birth cohort of 1,000 pairs of same-sex twins and their mothers followed prospectively through the twins' 18th year.

Objective 2. Elucidate mechanisms in brain and behavior that link polygenic risks with health problems. A key strength of the GWAS approach is that it can leapfrog current biological knowledge to make new discoveries. However, this necessitates follow-up research to uncover the mechanism through which genetic risk becomes manifest. In my postdoctoral research, I will pursue two collaborations to investigate candidate mechanisms through which genetic risk influences obesity and smoking.

(i) Taking polygenic risk inside the brain through imaging genetics research. One collaboration will be with brain researchers integrating genetics and neuroscience. Imaging genetics research seeks to understand the specific neural processes that connect genetic variation with differences in behavior.[237] Recent research in neuroscience links individual differences in neural phenotypes related to reward response, stress response, and attention and memory with obesity and with nicotine dependence and smoking cessation difficulties.[238-242] The collaboration with imaging genetics researchers will (1) identify specific neural phenotypes implicated in the pathogenesis of obesity and nicotine dependence; and (2) test associations between these phenotypes and the genetic risk scores. The goal of this collaboration will be to enhance understanding of the neurobiology linking genetic risk with health problems, with the aim of identifying specific treatment targets.

(ii) Taking polygenic risk into the field through genetically informed analysis of intervention trials. A second collaboration will be with investigators running intervention trials to address risk factors for smoking and obesity. Observational studies implicate behavioral and environmental pathways in the manifestation of genetic risk, e.g. sedentary lifestyle seem to be important in linking genetic risk with adult obesity and exposure to smoking cues seems to be important in linking genetic risks with smoking behavior.[185, 243] By utilizing intervention designs that explicitly manipulate behaviors and environmental triggers, we can better isolate the causal role of behaviors and environments in linking genetic risks with health problems. Depending on the size of the trials, it may also be possible to test whether individuals at higher genetic risk are

more or less susceptible to intervention effects. Collaboration with intervention researchers will (1) investigate whether the effects of genetic risk on obesity and smoking are altered by experimental manipulations of risk behaviors and environmental cues; and (2) depending on the size of the intervention cohorts, investigate whether genetic risks moderate treatment response.

Objective 3. Expand the developmental investigation of polygenic risk for obesity and smoking into the second half of the life course. My doctoral research characterized the relationship between genetic risks identified in GWAS and the development of obesity and nicotine dependence from childhood through mid-life. A next step in this research is to test how genetic risk relates to developmental phenotypes of obesity and nicotine dependence in the second half of the life course. My preliminary analyses using the ARIC cohort indicate that individuals at higher genetic risk lose weight more slowly and reduce their cigarette consumption more gradually in older age. Further, my preliminary analyses show that genetic risk for obesity predicts increased mortality risk.

Future research will first attempt to replicate the longitudinal associations between genetic risk and slower decline in body mass index and cigarette consumption in the second half of the life course. I will then move to address three new questions: (1) how do genetic risks for obesity and nicotine dependence relate to more general patterns of morbidity and to mortality risk in later life? (2) do health behavior changes following health shocks vary according to genetic risk? and (3) do genetic predispositions to obesity and nicotine dependence relate to health outcomes in different ways for individuals with different socioeconomic attainments.

APPENDIX A

Supplementary Materials to Chapter 2

This supplement describes the application of the 3-stage approach to create a genetic risk score (GRS) for obesity. The supplement is organized into 3 sections: The first section describes the creation of the obesity GRS: Stage 1. Extraction; Stage 2. Clustering; and Stage 3. Selection. The second section describes analyses comparing the resulting GRS to GRSs created with the best-guess and top-hits approaches. The final section describes sensitivity analyses to test heterogeneity in GRS associations.

PART 1. CREATING THE OBESITY GRS

Stage 1. Extraction

For our 3-stage approach analyses, we considered GWAS of European-descent samples that targeted 4 phenotypes: obesity, weight, waist circumference, and body mass index (BMI) (hereafter "obesity-related phenotypes"). A search of the NGHRI GWAS Catalog using the HuGE Navigator (http://www.hugenavigator.org) identified 16 GWAS that met these inclusion criteria, 9 of which were published by December 31, 2008 (**Supplementary Table 2.1**).

In Stage 1 (Extraction), we compiled association results reported in the manuscripts and supplementary materials of the GWAS and extracted rs-numbers and p-values for SNPs associated with any of the 4 phenotypes in the discovery or combined discovery and replication samples at an alpha level of $1x10^{-5}$ (n=103 SNPs in the subset of 9 GWAS, n=519 SNPs in the full set of 16 GWAS, **Supplementary Table 2.2**). The significance level of $p<1x10^{-5}$ was the most generous threshold at which most GWAS published results and is the threshold used in the NHGRI GWAS Catalog [3]. Associations were not extracted from replication samples because few GWAS reported novel associations identified in replication samples and some GWAS did not include replication samples or included replication samples of different ethnicity. Discovery sample risk SNPs that failed to replicate within an individual GWAS were included because replication was evaluated at the level of the GWAS publication rather than the specific test sample.

Stage 2. Clustering

In Stage 2 (Clustering), we grouped the extracted SNPs into "LD blocks." We defined LD blocks using data from the HapMap CEU sample (Phase 3), queried using Seattle SNPs' web-based Genome Variation Server (http://gvs.gs.washington.edu/GVS). For each SNP extracted in Stage 1 ("seeds"), we defined an LD block as the region containing all SNPs in LD with that seed at a threshold of $R^2 \geq 0.95$. Then, beginning with the block closest to the start of each chromosome, we pruned blocks that did not contain a unique seed. This process yielded n=66 LD blocks from the subset of 9 GWAS published by December 31, 2008 and n=158 LD blocks from the full set of 16 GWAS.

Stage 3. Selection

In Stage 3 (Selection), we retained LD blocks that we classified as genome-wide significant or as replicated. Genome-wide significant LD blocks were those that contained ≥1 SNP associated with an obesity-related phenotype at $p<1x10^{-8}$. Replicated blocks were those that contained SNPs extracted from ≥2 GWAS. This process yielded n=37 LD blocks clustered around 11 loci on chromosomes 1-4,9,11,12,16,18, and 19 from the subset of 9 GWAS and n=69 LD blocks clustered around 32 loci on chromosomes 1-6,9,11-14,16,18, and 19 from the full set of 16 GWAS (**Supplementary Tables 2.3, 2.4**). Sensitivity analyses relaxing the LD threshold used to define LD blocks yielded fewer LD blocks (e.g., for the full set of 16 GWAS, n=58 at an R^2 threshold of 0.70), but did not alter the loci identified as genome-wide significant or replicated in the original analyses.

PART 2. COMPARING THE 3-STAGE APPROACH GRSs TO THE TOP-HITS AND BEST-GUESS GRSs

To construct and test our GRSs, we followed-up the LD blocks identified in our 3-stage approach analyses in the GWAS dataset from the Atherosclerosis Risk in Communities (ARIC) Study. This dataset is publicly available through the National Institutes of Health Database of Genotypes and Phenotypes (dbGaP) (http://www.ncbi.nlm.nih.gov/gap, phs000090.v1.p1) and is described in the Data section of the main text.

We selected SNPs in the ARIC database to include in our two GRSs as follows: We defined tag SNPs for each of the LD blocks as SNPs that were in LD with every seed contained in the block at $R^2 \geq 0.95$. We then matched 1 tag SNP per LD block with a SNP in the ARIC study genotype database that met the GENEVA ARIC Project Team's quality control criteria [81]. If no tag SNPs in an LD block could be matched in the ARIC database, we relaxed the LD threshold used to define a tag SNP until either a) the resulting set of tag SNPs overlapped with tag SNPs that we had already matched in the ARIC database, or b) a match with a new SNP in the ARIC database was achieved. These analyses yielded a set of n=28 SNPs from the subset of 9 GWAS and a set of n=57 SNPs from the full set of 16 GWAS.

To compute the 3-stage approach GRSs for each ARIC participant, we (1) identified the obesity-associated allele for each SNP from the GWAS where that SNP was reported; (2) calculated the mean number of risk alleles at each locus; and (3) summed these means across loci to produce the 3-stage approach genome-wide scores.

To compute the top-hits and best-guess approach GRSs, we selected SNPs from the ARIC database to match SNPs from 3 published GRSs [61, 62, 89] and the full set of obesity-associated SNPs listed in the NHGRI GWAS catalog for GWAS of European-descent samples. In cases where a specific SNP was not available in the ARIC database, we selected its closest LD proxy. We then summed obesity-associated alleles across each set of selected SNPs to create the comparison genome-wide scores.

To test if the 3-stage approach could construct a GRS that was at least as predictive of BMI and obesity as GRSs created with the top-hits and best-guess approaches, we compared effect sizes for different GRSs using the ARIC data. All GRSs were standardized to have mean=0 and standard deviation=1. To measure GRS effect sizes for BMI, we estimated Pearson correlations (r) from separate linear regressions of BMI on each of the GRSs. To measure GRS effect sizes for obesity, we estimated odds ratios (OR) from separate logistic regressions of obesity on each of the GRSs. Regression models were adjusted for age (linear and quadratic terms), gender, the age-gender interaction, and the ARIC Study Centers where data were collected (hereafter these statistical adjustments are described as "demographics and geography"). To test differences between GRS effect sizes, we conducted F-tests (for effect sizes estimated from linear regressions) and Wald tests (for effect sizes estimated from logistic

regressions). For these tests, models including each of the GRSs being compared were jointly estimated using the seemingly unrelated regression method. Seemingly unrelated regression is a statistical approach for comparing coefficients from non-nested regression models [111, 112]. Effect sizes were similar for all GRSs. Statistical tests indicated that our 3-stage approach GRSs performed as well as or better than GRSs created using top-hits and best-guess approaches (**Supplementary Table 2.5**). Thus, the 3-stage approach produced a GRS that was at least as predictive as top-hits and best guess approach GRSs. We used the 3-stage approach GRS created from the full set of 16 GWAS (hereafter the "Obesity GRS") in subsequent analyses.

Refining the 3-Stage Approach GRS for Obesity. At 7 of the 32 loci identified in the 3-stage approach analyses of GWAS results (in or near the genes *TMEM18*, *ETV5*, *BDNF*, *MTCH2*, *FTO*, *MC4R*, and *KCTD15*), multiple LD blocks met selection criteria (genome-wide significance or replication). To refine the 3-stage approach GRS, we asked whether the genotype for a single SNP could be used instead of the mean number of risk alleles at a locus. First, we identified the BMI-increasing allele for each SNP and calculated the linear association between the number of BMI-increasing alleles for that SNP and BMI measured at the first ARIC study visit. We next compared test statistics and effect sizes between SNPs at each locus to identify the "lead-SNP", the SNP with the strongest association, and the worst-associated SNP. We then compared the effect size for the lead-SNP to the effect sizes for the worst-associated SNP and for the mean number of risk alleles across SNPs at the locus. These analyses asked 1) whether there was any difference in the signal from the different SNPs in a correlated set; and 2) whether a single SNP could provide an adequate summary of obesity-associated variation at the locus. Models were fitted using linear regression with statistical adjustment for demographics and geography. We compared effect sizes using the seemingly unrelated regression method [111, 112]. **Supplementary Table 2.6** shows results from this analysis. At all loci, the lead SNP, worst-associated SNP, and mean number of risk alleles performed similarly, with the exception of the *FTO* locus, at which the lead SNP rs9939609 performed slightly better than the worst-associated SNP rs1477196. Finally, we tested whether including multiple SNPs at a locus improved the prediction of BMI in a regression model. Analyses were conducted using the variable selection algorithm in the Stata program mfp [113]. Details of this method are reported elsewhere [114]. Briefly, SNPs were added to a baseline model predicting BMI as a function of age, sex, and geography in order of decreasing statistical significance of the SNPs' bivariate association with BMI. SNPs were retained in the

model if their inclusion resulted in a statistically significant ($p<0.05$) decrease in model deviance. Results showed that model fit was not improved by the inclusion of multiple SNPs at any locus. Therefore, we retained only the best-associated SNPs from each of the 7 loci, resulting in a 32-SNP GRS (**Supplementary Table 2.7**).

PART 3. SENSITIVITY ANALYSES TO TEST HETEROGENEITY IN GRS ASSOCIATIONS

We tested the linearity of GRS-BMI associations using quadratic and cubic specifications of the GRS in linear regression models. Coefficients for the higher order (i.e. squared and cubic) GRS terms were not statistically significant ($p>0.10$ for all), indicating that the GRS-BMI association was approximately linear. We tested the measurement specificity of GRS-BMI associations by comparing GRS effect sizes for BMI to GRS effect sizes for weight and for waist circumference using the seemingly unrelated regression method [112]. GRS coefficients were similar across all three models ($p>0.10$ for tests of differences), indicating that the GRS predicted not just BMI, but related measures of body size and adiposity. We tested the whether GRS-BMI associations were different for men and women or for older as compared to younger individuals using product terms in linear regression models. Coefficients for product terms were not statistically significant ($p>0.10$ for all), indicating that GRS-BMI associations were similar for men and women and for older and younger individuals. Finally, we tested whether GRS-BMI associations differed across the 4 in-person assessments in the ARIC Study using the seemingly unrelated regression method. GRS effect sizes were similar across all 4 assessments ($p>0.10$ for all comparisons), indicating that GRS-BMI associations were consistent across measurement intervals.

Supplementary Table 2.1. Genome Wide Association Studies Included In 3-Stage Approach Analyses. GWAS information comes from the NHGRI GWAS Catalog (www.genome.gov). Risk SNPs were defined as any SNP associated with an obesity-related phenotype (BMI, weight, waist circumference, categorical obesity) at $p<10^{-5}$ in the discovery or combined discovery and replication samples of the GWAS. *Italicized counts include imputed genotypes; **Lindgren et al. also investigated associations with waist circumference, and these are the association tests included in the SNP selection analysis; ***Scherag et al. also investigated associations with BMI and both phenotypes were included in the SNP selection analysis. Citations for the GWAS are included as [89, 90, 115-129].

	GWAS Chip Manufacturer	SNPs Genotyped*	SNPs in GWAS Catalog		Risk SNPs Included in Analyses
			SNPs	Phenotypes	
Herbert et al. 2006	Affymetrix	86,604	0	Obesity	0
Frayling et al. 2007	Affymetrix	490,032	1	BMI	1
Scuteri et al. 2007	Affymetrix	362,129	1	BMI, Weight	12
Fox et al. 2007	Affymetrix	70,897	5	BMI, Waist Circumference	12
Hinney et al. 2007	Affymetrix	440,794	1	Obesity (early onset extreme)	15
Liu et al. 2008	Affymetrix	379,319	0	Obesity	3
Loos et al. 2008	Affymetrix	344,883	2	BMI	10
Thorleifsson et al. 2009	Illumina	305,846	18	BMI, Weight	47
Willer et al. 2009	Affymetrix & Illumina	*2,399,588*	11	BMI	24
Meyre et al. 2009	Illumina	308,846	5	Obesity	32
Cotsapas et al. 2009	Illumina	457,251	13	Obesity (extreme)	15
Lindgren et al. 2009	Affymetrix & Illumina	2,573,738	NA	Adiposity**	10
Heard-Costa et al. 2009	Affymetrix & Illumina	512,349	7	Waist Circumference	320
Johansson et al. 2009	Illumina	318,237	17	BMI, Weight	26
Liu et al. 2010	Illumina	*559,712*	2	BMI	3
Scherag et al. 2010	Affymetrix & Illumina	*1,596,878*	2	Obesity (extreme)***	13
Speliotes et al. 2010	Affymetrix, Illumina, Perlegen	*~2.8 million*	38	BMI	42

Supplementary Table 2.2. Risk SNPs and Source Publications: All SNPs reported as associated with Obesity, BMI, Weight, or Waist Circumference at $p<1x10^{-5}$ in Discovery or Combined Discovery and Replication Samples

Risk SNP	Trait	Publication
rs9939609	BMI	Frayling et al. 2007 Science
rs1121980	BMI	Scuteri et al. 2007
rs6602024	BMI	
rs7193144	BMI	
rs8050136	BMI	
rs9926289	BMI	
rs9930506	BMI	
rs9939609	BMI	
rs9939973	BMI	
rs9940128	BMI	
*rs4512445**	Waist Circumference	
rs7193144	Waist Circumference	
rs8050136	Waist Circumference	
rs1106683	BMI	Fox et al. 2007
rs1106684	BMI	
rs1333026	BMI	
rs10488165	Waist Circumference	
rs10504576	Waist Circumference	
rs1875517	Waist Circumference	
rs2206682	Waist Circumference	
rs2223662	Waist Circumference	
rs4469448	Waist Circumference	
rs4471028	Waist Circumference	
rs6996971	Waist Circumference	
rs953536	Waist Circumference	
rs10008032	Extreme Obesity	Hinney et al. 2007
rs1121980	Extreme Obesity	
rs16998603	Extreme Obesity	
rs2172478	Extreme Obesity	
rs2969001	Extreme Obesity	
rs3783950	Extreme Obesity	
rs41492957	Extreme Obesity	
rs6076920	Extreme Obesity	
rs619819	Extreme Obesity	
rs7193144	Extreme Obesity	
rs8050136	Extreme Obesity	
rs9276431	Extreme Obesity	
rs9939609	Extreme Obesity	
rs9939973	Extreme Obesity	
rs9940128	Extreme Obesity	

Supplementary Table 2 Continued		
Risk SNP	Trait	Publication
rs16986921	BMI	
rs6013029	BMI	
rs6020712	BMI	Liu et al. 2008
rs10498767	BMI	
rs1121980	BMI	
rs17700633	BMI	
rs17782313	BMI	
rs2572106	BMI	Loos et al. 2008
rs2679120	BMI	
rs4623795	BMI	
rs7212681	BMI	
rs7336049	BMI	
rs748192	BMI	
rs10501087	BMI	
rs10783050	BMI	
rs10913469	BMI	
rs12970134	BMI	
rs1776012	BMI	
rs2568958	BMI	
rs2867125	BMI	
rs29941	BMI	
rs3101336	BMI	
rs3751812	BMI	
rs4074134	BMI	
rs467650	BMI	
rs4788102	BMI	
rs4854344	BMI	
rs4923461	BMI	
rs6265	BMI	
rs6499640	BMI	
rs7138803	BMI	
rs7190492	BMI	
rs7336332	BMI	
rs7481311	BMI	
rs7498665	BMI	
rs7561317	BMI	
rs7647305	BMI	Thorleifsson et al. 2009
rs7647305	BMI	
rs8044769	BMI	
rs8049439	BMI	
rs8050136	BMI	
rs836964	BMI	
rs867559	BMI	
rs925946	BMI	
rs9424977	BMI	
rs1047440	Weight	
rs1077393	Weight	
rs10835211	Weight	
rs1350341	Weight	
rs1350341	Weight	
rs17069257	Weight	
rs1973993	Weight	
rs2115172	Weight	
rs2260000	Weight	
rs2260000	Weight	
rs2844479	Weight	
rs2844479	Weight	
rs3766431	Weight	
rs633265	Weight	
rs6477693	Weight	

Supplementary Table 2 Continued		
Risk SNP	Trait	Publication
rs10769908	BMI	
rs10769908	BMI	
rs10838738	BMI	
rs10838738	BMI	
rs10938397	BMI	
rs10938397	BMI	
rs11084753	BMI	
rs11084753	BMI	
rs11084753	BMI	
rs11773921	BMI	
rs12324805	BMI	
rs1421085	BMI	
rs1439845	BMI	
rs17700144	BMI	
rs17782313	BMI	
rs17782313	BMI	
rs2145270	BMI	Willer et al. 2009
rs2145270	BMI	
rs2245715	BMI	
rs2815752	BMI	
rs2815752	BMI	
rs2815752	BMI	
rs4752856	BMI	
rs6548238	BMI	
rs6548238	BMI	
rs6548238	BMI	
rs6907460	BMI	
rs7181095	BMI	
rs7498665	BMI	
rs7498665	BMI	
rs752238	BMI	
rs9931989	BMI	
rs9939609	BMI	
rs9939609	BMI	
rs10508503	Obesity	
rs11071927	Obesity	
rs11956401	Obesity	
rs12588659	Obesity	
rs12633433	Obesity	
rs1326986	Obesity	
rs1343772	Obesity	
rs1380100	Obesity	
rs1396618	Obesity	
rs1421085	Obesity	
rs1424233	Obesity	
rs16829231	Obesity	
rs17782313	Obesity	
rs1805081	Obesity	
rs1858367	Obesity	
rs2011946	Obesity	Meyere et al. 2009
rs2158044	Obesity	
rs2908338	Obesity	
rs3026762	Obesity	
rs3102841	Obesity	
rs413693	Obesity	
rs4712652	Obesity	
rs4786847	Obesity	
rs6463923	Obesity	
rs646839	Obesity	
rs6580742	Obesity	
rs6796959	Obesity	
rs7506051	Obesity	
rs7717673	Obesity	
rs908078	Obesity	
rs9275582	Obesity	
rs987052	Obesity	

Supplementary Table 2 Continued

Risk SNP	Trait	Publication
rs10433903	Extreme Obesity	
rs10999409	Extreme Obesity	
rs12295638	Extreme Obesity	
rs12492816	Extreme Obesity	
rs12635698	Extreme Obesity	
rs1435703	Extreme Obesity	
rs2274459	Extreme Obesity	
rs374748	Extreme Obesity	Cotsapas et al. 2009
rs6110577	Extreme Obesity	
rs6726292	Extreme Obesity	
rs7474896	Extreme Obesity	
rs7603514	Extreme Obesity	
rs9366829	Extreme Obesity	
rs9941349	Extreme Obesity	
rs999943	Extreme Obesity	
rs10085177	Waist Circumference	
rs11970116	Waist Circumference	
rs13116494	Waist Circumference	
rs2245667	Waist Circumference	
rs4737325	Waist Circumference	Lindgren et al. 2009
rs6429082	Waist Circumference	
rs7194591	Waist Circumference	
rs7826222	Waist Circumference	
rs7970350	Waist Circumference	
rs987237	Waist Circumference	
rs10096750	BMI	
rs10145154	BMI	
rs10146997	BMI	
rs10150332	BMI	
rs10173167	BMI	
rs10188334	BMI	
rs10189761	BMI	
rs10190052	BMI	
rs10193244	BMI	
rs10511835	BMI	
rs10813208	BMI	
rs10852521	BMI	
rs10871777	BMI	
rs10875982	BMI	
rs10969478	BMI	
rs11075985	BMI	
rs11075987	BMI	
rs11075989	BMI	
rs11075990	BMI	
rs11127483	BMI	
rs11127484	BMI	
rs11127485	BMI	Heard-Costa et al. 2009
rs11127491	BMI	
rs11152213	BMI	
rs11169176	BMI	
rs1121980	BMI	
rs11520442	BMI	
rs11642841	BMI	
rs11660783	BMI	
rs11662368	BMI	
rs11663816	BMI	
rs11664883	BMI	
rs11665563	BMI	
rs12002080	BMI	
rs12149832	BMI	
rs12446228	BMI	
rs12623218	BMI	
rs12714414	BMI	
rs12714415	BMI	
rs12954782	BMI	
rs12955983	BMI	
rs12957347	BMI	
rs12960928	BMI	
rs12964203	BMI	

Supplementary Table 2 Continued		
Risk SNP	Trait	Publication
rs12966550	BMI	
rs12967135	BMI	
rs12969709	BMI	
rs12970134	BMI	
rs12992154	BMI	
rs12995480	BMI	
rs13007080	BMI	
rs13007086	BMI	
rs13012571	BMI	
rs13021737	BMI	
rs1320330	BMI	
rs1320331	BMI	
rs1320336	BMI	
rs1320337	BMI	
rs1320338	BMI	
rs13386517	BMI	
rs13386627	BMI	
rs13386964	BMI	
rs13388043	BMI	
rs13393304	BMI	
rs13396935	BMI	
rs13397165	BMI	
rs13401686	BMI	
rs13415094	BMI	
rs1350341	BMI	
rs1421085	BMI	
rs1456404	BMI	
rs1457489	BMI	
rs1477196	BMI	
rs1539952	BMI	
rs1553754	BMI	
rs1555967	BMI	
rs1558902	BMI	
rs1619975	BMI	
rs1673518	BMI	Heard-Costa et al. 2009
rs17109256	BMI	
rs17175643	BMI	
rs17201502	BMI	
rs17299673	BMI	
rs17700144	BMI	
rs17782313	BMI	
rs17817288	BMI	
rs17817449	BMI	
rs17817964	BMI	
rs1861866	BMI	
rs1861867	BMI	
rs1942860	BMI	
rs1942863	BMI	
rs1942866	BMI	
rs2051311	BMI	
rs2051312	BMI	
rs2058908	BMI	
rs2168708	BMI	
rs2168711	BMI	
rs2206277	BMI	
rs2288278	BMI	
rs2331841	BMI	
rs2397026	BMI	
rs2860323	BMI	
rs2867108	BMI	
rs2867109	BMI	
rs2867110	BMI	
rs2867112	BMI	
rs2867113	BMI	
rs2867122	BMI	
rs2867123	BMI	
rs2867125	BMI	
rs2867131	BMI	
rs2903492	BMI	

Supplementary Table 2 Continued

Risk SNP	Trait	Publication
rs2947411	BMI	
rs297924	BMI	
rs34341	BMI	
rs3751812	BMI	
rs3751813	BMI	
rs3928247	BMI	
rs4045166	BMI	
rs4299252	BMI	
rs4423631	BMI	
rs4438957	BMI	
rs4452188	BMI	
rs4613321	BMI	
rs4615388	BMI	
rs4620360	BMI	
rs474112	BMI	
rs475134	BMI	
rs476828	BMI	
rs4783819	BMI	
rs4784323	BMI	
rs4793927	BMI	
rs4854344	BMI	
rs4854348	BMI	
rs4854349	BMI	
rs487720	BMI	
rs489693	BMI	
rs492443	BMI	
rs497353	BMI	
rs5017300	BMI	
rs5017303	BMI	
rs521663	BMI	
rs523288	BMI	
rs536783	BMI	
rs538656	BMI	
rs545708	BMI	
rs559623	BMI	Heard-Costa et al. 2009
rs562622	BMI	
rs563726	BMI	
rs565239	BMI	
rs565970	BMI	
rs571312	BMI	
rs574988	BMI	
rs589850	BMI	
rs590215	BMI	
rs591166	BMI	
rs611428	BMI	
rs633265	BMI	
rs649721	BMI	
rs6499640	BMI	
rs6548237	BMI	
rs6567155	BMI	
rs6567160	BMI	
rs6567161	BMI	
rs663129	BMI	
rs666181	BMI	
rs6711012	BMI	
rs6719518	BMI	
rs6719980	BMI	
rs6725549	BMI	
rs6728726	BMI	
rs6731348	BMI	
rs6731688	BMI	
rs6732471	BMI	
rs6734363	BMI	
rs6742576	BMI	
rs6743060	BMI	
rs6744646	BMI	
rs6744653	BMI	
rs6745266	BMI	
rs6752470	BMI	

Supplementary Table 2 Continued		
Risk SNP	Trait	Publication
rs6755502	BMI	
rs681630	BMI	
rs682614	BMI	
rs683430	BMI	
rs7022642	BMI	
rs7132908	BMI	
rs7138803	BMI	
rs7144011	BMI	
rs7185735	BMI	
rs7190492	BMI	
rs7193144	BMI	
rs7201850	BMI	
rs7202116	BMI	
rs7203521	BMI	
rs7205986	BMI	
rs7206010	BMI	
rs7206790	BMI	
rs7240566	BMI	
rs7338657	BMI	
rs7561317	BMI	
rs7567570	BMI	
rs7570198	BMI	
rs7571957	BMI	
rs7574359	BMI	
rs7576624	BMI	
rs7576635	BMI	
rs7585056	BMI	
rs7587786	BMI	
rs7604609	BMI	
rs7608050	BMI	
rs7715806	BMI	
rs7831920	BMI	
rs8043757	BMI	Heard-Costa et al.
rs8044769	BMI	2009
rs8047395	BMI	
rs8050136	BMI	
rs8051591	BMI	
rs8055197	BMI	
rs8057044	BMI	
rs8083289	BMI	
rs8086627	BMI	
rs8089364	BMI	
rs8091524	BMI	
rs8095404	BMI	
rs921971	BMI	
rs939582	BMI	
rs939583	BMI	
rs953442	BMI	
rs975918	BMI	
rs981106	BMI	
rs981113	BMI	
rs987237	BMI	
rs9922047	BMI	
rs9922619	BMI	
rs9922708	BMI	
rs9923147	BMI	
rs9923233	BMI	
rs9923544	BMI	
rs9928094	BMI	
rs9930333	BMI	
rs9930501	BMI	
rs9930506	BMI	
rs9931494	BMI	
rs9932754	BMI	
rs9935401	BMI	
rs9936385	BMI	
rs9937053	BMI	

Supplementary Table 2 Continued

Risk SNP	Trait	Publication
rs993887	BMI	
rs9939609	BMI	
rs9939973	BMI	
rs9940128	BMI	
rs9940646	BMI	
rs9941349	BMI	
rs10059683	Waist Circumference	
rs10066756	Waist Circumference	
rs10068332	Waist Circumference	
rs10146690	Waist Circumference	
rs10150482	Waist Circumference	
rs10869557	Waist Circumference	
rs10869558	Waist Circumference	
rs10869559	Waist Circumference	
rs11778132	Waist Circumference	
rs11780082	Waist Circumference	
rs11857639	Waist Circumference	
rs11990688	Waist Circumference	
rs12271537	Waist Circumference	
rs12274672	Waist Circumference	
rs12475139	Waist Circumference	
rs12792768	Waist Circumference	
rs13404551	Waist Circumference	
rs1447905	Waist Circumference	
rs1521252	Waist Circumference	
rs16930931	Waist Circumference	
rs17008958	Waist Circumference	
rs17061143	Waist Circumference	
rs17109221	Waist Circumference	
rs17476669	Waist Circumference	
rs17537900	Waist Circumference	
rs17836088	Waist Circumference	
rs2164210	Waist Circumference	
rs2236783	Waist Circumference	
rs2322659	Waist Circumference	Heard-Costa et al.
rs2322660	Waist Circumference	2009
rs2365642	Waist Circumference	
rs2370982	Waist Circumference	
rs303211	Waist Circumference	
rs309134	Waist Circumference	
rs309137	Waist Circumference	
rs309160	Waist Circumference	
rs309168	Waist Circumference	
rs4098360	Waist Circumference	
rs4420638	Waist Circumference	
rs4701252	Waist Circumference	
rs4758213	Waist Circumference	
rs4758215	Waist Circumference	
rs507824	Waist Circumference	
rs569406	Waist Circumference	
rs6499641	Waist Circumference	
rs6714750	Waist Circumference	
rs6716536	Waist Circumference	
rs6754311	Waist Circumference	
rs6817633	Waist Circumference	
rs6837818	Waist Circumference	
rs6870971	Waist Circumference	
rs687670	Waist Circumference	
rs693895	Waist Circumference	
rs6998794	Waist Circumference	
rs7110070	Waist Circumference	
rs7156625	Waist Circumference	
rs745500	Waist Circumference	
rs748841	Waist Circumference	
rs7579771	Waist Circumference	
rs7824886	Waist Circumference	
rs7932813	Waist Circumference	
rs8059991	Waist Circumference	
rs892715	Waist Circumference	
rs9598518	Waist Circumference	
rs9790104	Waist Circumference	

Supplementary Table 2 Continued

Risk SNP	Trait	Publication
rs1024889	BMI	Johansson et al. 2009
rs1152846	BMI	
rs12517906	BMI	
rs1458095	BMI	
rs1878047	BMI	
rs1927702	BMI	
rs2383393	BMI	
rs3803915	BMI	
rs3803915	BMI	
rs3934834	BMI	
rs4085400	BMI	
rs824931	BMI	
rs875283	BMI	
rs10844154	Weight	
rs10972341	Weight	
rs10972350	Weight	
rs1152846	Weight	
rs12517906	Weight	
rs1570885	Weight	
rs1816002	Weight	
rs1840440	Weight	
rs2765086	Weight	
rs4879869	Weight	
rs7209395	Weight	
rs7919006	Weight	
rs965178	Weight	
rs2275215	BMI	Liu et al. 2010
rs10458787	BMI	
rs11127485	BMI	Scherag et al. 2010
rs1558902	BMI	
rs9935401	BMI	
rs10926984	Obesity	
rs12145833	Obesity	
rs2783963	Obesity	
rs11127485	Obesity	
rs17150703	Obesity	
rs13278851	Obesity	
rs516175	Obesity	
rs1558902	Obesity	
rs9935401	Obesity	
rs17700144**	Obesity	

Supplementary Table 2 Continued		
Risk SNP	Trait	Publication
rs1558902	BMI	
rs2860323	BMI	
rs6567160	BMI	
rs10938397	BMI	
rs10767664	BMI	
rs543874	BMI	
rs2815752	BMI	
rs10182181	BMI	
rs12444979	BMI	
rs7498665	BMI	
rs987237	BMI	
rs2241423	BMI	
rs9816226	BMI	
rs7138803	BMI	
rs2287019	BMI	
rs1514177	BMI	
rs13107325	BMI	
rs2112347	BMI	
rs10968576	BMI	
rs3817334	BMI	
rs3810291	BMI	Speliotes et al. 2010
rs887912	BMI	
rs10150332	BMI	
rs7640855	BMI	
rs11847697	BMI	
rs2890652	BMI	
rs11165643	BMI	
rs4771122	BMI	
rs4836133	BMI	
rs4929949	BMI	
rs29938	BMI	
rs9296115	BMI	
rs2922763	BMI	
rs2444217	BMI	
rs867559	BMI	
rs3764400	BMI	
rs255414	BMI	
rs6955651	BMI	
rs17016663	BMI	
rs6477694	BMI	
rs2652594	BMI	
rs2035935	BMI	

Supplementary Table 2.2 Footnote: *Reported as "SNP_A-2284869" and crosswalked to rs ID using the Affy 6.0 SNP name to rs ID crosswalk file "GenomeWideSNP_6.na30.annot.csv"; **The GWAS catalog reports rs10871777 (in LD with rs17700144 at R^2=0.85) as the obesity-associated SNP near the gene MC4R in Scherag et al. SNPs are reported only once per GWAS. Associations are reported for BMI where present and for other phenotypes where BMI was not investigated or the SNP was not associated with BMI at $p<1 \times 10^{-5}$

Supplementary Table 2.3. Replicated and/or Genome-Wide Significant LD Blocks Identified in 3-Stage Approach Analyses. LD blocks were defined from LD analyses of risk SNPs (genotype-phenotype association at $p<1x10^{-5}$) using data from the HapMap version 3 CEU sample accessed via Seattle SNPs's Genome Variation Server and an LD threshold of $R^2 \geq 0.95$. Replication was evaluated as the number of GWAS reporting any SNP in the block as a risk SNP. Genes were evaluated within 100kb in either direction from an LD block's outermost SNPs.

Chromsome	Identified LD Blocks	Replicated LD Blocks	Mean Number of Replications (All Blocks)	Genes
1	4	3	2.0	*NEGR1, TNNI3K, PTB2, SEC16B*
2	6	2	2.0	*LRP1B, TMEM18*
3	3	0	1.0	*CADM2, ETV5/DGKG*
4	2	1	1.5	*GNPDA2, SLC39A8*
5	2	0	1.0	*POC5, ZNF608*
6	1	1	3.0	*TFAP2B*
9	2	1	1.5	*LING02/LRRN6C, LMX1B*
11	7	0	1.0	*RPL27A, BDNF, MTCH2*
12	1	1	3.0	*BDCDIN3D/FAIM2/NCKAP5L*
13	1	0	3.0	*MTIF3, GRF3A*
14	2	1	1.5	*PRKD1, NRXN3*
15	1	0	1.0	*MAP2K5*
16	26	14	3.0	*GRP5B, ATXN2L/TUFM/SH2B1, FTO*
18	7	7	2.6	*MC4R*
19	4	1	1.3	*KCTD15, ZC3H4, QPCTL, TMEM160*

Supplementary Table 2.4. Characteristics of Replicated and/or Genome-Wide Significant LD Blocks

	LD Block				GWAS Publication														
Chrom-osome	Chromosomal Space Covered by All Risk SNPs in the LD Block (NCBI Build 36)	Nearby Genes	Seed SNPs (risk SNPs in LD with all risk SNPs in block at R^2≥0.95) // Proxy SNPs (risk SNPs in LD with any seed SNP at R2≥0.95)	Any SNP in Block Genome-Wide Significant	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]
1	72,523,773 - 72,585,028	NEGR1	rs2568958, rs2815752, rs3101336	Yes							X	X								X
	74,763,990	TNNI3K	rs1514177	Yes																X
	96,696,685 - 96,716,582	*PTBP2*	rs11165643 // rs1973993	Yes							X									X
	176,156,103 - 176,180,142	SEC16B	rs10913469, rs543874	Yes							X									X
2	604,168 - 643,874	TMEM18	See footnote	Yes							X					X			X	X
	604,210 - 643,874	TMEM18		Yes							X					X			X	X
	624,905	TMEM18	rs6548238	Yes								X								
	25,003,800		rs10182181	Yes																X
	59,156,381		rs887912	Yes																X
	142,676,401	LRP1B	rs2890652	Yes																X
3	85,956,854	CADM2	rs7640855	Yes																X
	187,316,984	ETV5/DGKG	rs7647305	Yes							X									
	187,317,193	ETV5/DGKG	rs9816226	Yes																X
4	44,877,284		rs10938397	Yes								X								X
	103,407,732	SLC39A8	rs13107325	Yes																X
5	75,050,998	POC5	rs2112347	Yes																X
	124,360,002		rs4836133	Yes																X
6	50,906,485 - 50,911,009	TFAP2B	rs2206277, rs987237	Yes											X	X				X
9	28,404,339	LING02	rs10968576	Yes																X
	128,505,146	LMX1B	rs867559	$p<1x10^{-6}$							X									X
11	8,561,169	STK33	rs4929949	Yes																X
	27,603,861 - 27,626,684	BDNF	rs10501087, rs4074134, rs4923461	Yes							X									
	27,636,492	BDNF	rs6265	Yes							X									
	27,682,562	BDNF	rs10767664	Yes																X
	27,623,778 - 27,623,778	BDNF	rs925946	Yes							X									
	47,604,618 - 47,619,625	MTCH2	rs10838738, rs4752856	Yes								X								
	47,607,569	MTCH2	rs3817334	Yes																X
12	48,533,735	BDCDIN3D, FAIM2, NCKAP5L	rs7138803	Yes							X					X				X
13	26,918,180	MTIF3, *GRF3A*	rs4771122	Yes																X
14	29,584,863		rs11847697	Yes																X
	78,961,635 - 79,014,915	NRXN3	rs10145154, rs10150332, rs17109256, rs7144011 // rs10146997, rs10150482, rs17109221, rs17836088, rs7156625	Yes												X				X
15	65,873,892	MAP2K5	rs2241423	Yes																X

Supplementary Table 4 Continued

			LD Block		GWAS Publication														
Chrom-osome	Chromosomal Space Covered by All Risk SNPs in the LD Block (NCBI Build 36)	Genes Overlapping LD Block/ 10kb of SNP*	Seed SNPs (risk SNPs in LD with all risk SNPs in block at R^2≥0.95) // Proxy SNPs (risk SNPs in LD with any seed SNP at R2≥0.95)	Any SNP in Block Genome-Wide Significant	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]
16	19,841,101	GPRC5B	rs12444979	Yes																X
	28,745,016 - 28,790,742	ATXN2L, TUFM, SH2B1	rs4788102, rs7498665, rs8049439	Yes							X	X								X
	52,312,678 - 52,327,178	FTO	rs6499640, rs7203521, rs7206010	Yes							X					X				
	52,355,409	FTO	rs7206790	Yes												X				
	52,356,024 - 52,361,841	FTO	rs8047395 //rs1861866, rs8055197	Yes												X				
	52,356,024 - 52,363,781	FTO	rs1861866, rs8055197 // rs10852521, rs8047395, rs9922047	Yes												X				
	52,360,657 - 52,372,662	FTO	rs10852521, rs9922047 // rs11075987, rs1861866, rs8055197	Yes												X				
	52,362,466 - 52,372,662	FTO	rs11075987 // rs10852521, rs9922047	Yes												X				
	52,365,265	FTO	rs17817288	Yes												X				
	52,370,115	FTO	rs8057044	Yes												X				
	52,396,636	FTO	rs8044769	Yes							X					X				
	52,357,008 - 52,366,748	FTO	rs11075985, rs9940646 // rs1121980, rs9923147, rs9923544, rs9928094, rs9930333, rs9937053, rs9939973, rs9940128	Yes		X		X		X						X				
	52,357,008 - 52,384,680	FTO	rs1121980, rs9923147, rs9923544, rs9928094, rs9930333, rs9937053, rs9939973, rs9940128 // rs11075985, rs1421085, rs1558902, rs7201850, rs9931494, rs9940646, rs9941349	Yes		X		X		X		X	X	X		X			X	X
	52,357,008 - 52,385,567	FTO	rs1421085, rs1558902 // rs17817964, rs7185735, rs7193144, rs7202116, rs9937053	Yes		X		X				X	X			X			X	X
	52,357,008 - 52,389,272	FTO	rs7201850, rs9931494, rs9941349 // rs1121980, rs9922619, rs9922708, rs9923147, rs9923544, rs9928094, rs9930333, rs9930501, rs9930506, rs9932754, rs9937053, rs9939973, rs9940128	Yes		X		X		X				X		X				
	52,358,455 - 52,400,409	FTO	rs17817964, rs7185735 // rs11075989, rs11075990, rs12149832, rs1421085, rs1558902, rs17817449, rs3751812, rs7193144, rs7202116, rs8043757, rs8050136, rs8051591, rs9923233, rs9935401, rs9939609	Yes	X	X		X			X	X	X			X			X	X
	52,361,075 - 52,400,409	FTO	rs7193144, rs7202116 // rs11075989, rs11075990, rs12149832, rs1558902, rs17817449, rs17817964, rs3751812, rs7185735, rs8043757, rs8050136, rs8051591, rs9923233, rs9935401, rs9939609	Yes	X	X		X			X	X				X			X	X
	52,368,187 - 52,385,567	FTO	rs11075989, rs11075990, rs17817449, rs3751812, rs8043757, rs8050136, rs8051591, rs9923233, rs9935401, rs9939609 // rs17817964, rs7185735, rs7193144, rs7202116, rs9936385	Yes	X	X		X			X	X				X			X	
	52,368,187 - 52,400,409	FTO	rs12149832 // rs17817964, rs7185735, rs7193144, rs7202116	Yes		X		X								X				
	52,376,670 - 52,377,378	FTO	rs9936385 // rs11075989, rs9923233	Yes												X				
	52,379,363 - 52,389,272	FTO	rs9922619, rs9922708, rs9930501, rs9932754 // rs7201850, rs9930506, rs9931494, rs9941349	Yes		X								X		X				
	52,382,989 - 52,389,272	FTO	rs9930506 // rs9922619, rs9922708, rs9930501, rs9931494, rs9932754, rs9941349	Yes		X								X		X				
	52,406,062	FTO	rs1861867	Yes												X				
	52,357,888 - 52,386,253	FTO	rs12446228, rs1477196, rs4783819, rs7190492	Yes							X					X				
	52,376,209	FTO	rs3751813	Yes												X				
	52,402,988	FTO	rs11642841	Yes												X				

Supplementary Table 4 Continued

	LD Block				GWAS Publication														
Chrom-osome	Chromosomal Space Covered by All Risk SNPs in the LD Block (NCBI Build 36)	Genes Overlapping LD Block/ 10kb of SNP*	Seed SNPs (risk SNPs in LD with all risk SNPs in block at R^2≥0.95) // Proxy SNPs (risk SNPs in LD with any seed SNP at R2≥0.95)	Any SNP in Block Genome-Wide Significant	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]
18	55,962,962	*MC4R*	rs17700144	$p<1x10^{-6}$								X				X			X	
	55,980,115 - 56,003,928	*MC4R*	rs10871777, rs11152213, rs12967135, rs17782313, rs2168711, rs476828, rs523288, rs538656, rs571312, rs6567160, rs663129	Yes						X		X	X			X				X
	55,964,628 - 56,003,732	*MC4R*	rs1350341, rs1619975, rs1673518, rs2051311, rs2051312, rs2331841, rs474112, rs475134, rs487720, rs536783, rs545708, rs559623, rs562622, rs565239, rs565970, rs574988, rs589850, rs591166, rs611428, rs649721, rs6567161, rs666181, rs681630, rs682614, rs683430, rs975918, rs993887 // rs521663, rs633265	$p<1x10^{-6}$								X				X				
	56,009,782 - 56,048,783	*MC4R*	rs12960928 // rs11663816, rs11664883, rs11665563, rs12954782, rs12969709, rs12970134, rs1457489, rs17175643, rs492443, rs8083289, rs8089364, rs921971	Yes								X				X				
	56,009,782 - 56,062,310	*MC4R*	rs921971 // rs11663816, rs11664883, rs11665563, rs12954782, rs12955983, rs12960928, rs12964203, rs12966550, rs12969709, rs12970134, rs1457489, rs17175643, rs2168708, rs492443, rs8083289, rs8089364	Yes								X				X				
	56,009,809 - 56,047,722	*MC4R*	rs12955983 // rs11663816, rs11664883, rs11665563, rs12954782, rs12969709, rs12970134, rs1457489, rs17175643, rs8083289, rs8089364, rs921971	Yes								X				X				
	56,009,809 - 56,062,310	*MC4R*	rs11663816, rs11664883, rs11665563, rs12954782, rs12964203, rs12966550, rs12969709, rs12970134, rs1457489, rs17175643, rs2168708, rs8083289, rs8089364 // rs12955983, rs12960928, rs921971	Yes								X				X				
19	39,001,372 - 39,003,321	KCTD15	rs29938, rs29941	Yes								X								X
	39,013,977	*KCTD15*	rs11084753	Yes									X							
	52,260,843	ZC3H4, TMEM160	rs3810291	Yes																X
	50,894,012	QPCTL	rs2287019	Yes																X

Supplementary Table 2.4 Footnote: GWAS are numbered as follows: [1] Frayling et al. 2007, *Science*; [2] Scuteri et al. 2007, *PLoS Genetics*; [3] Fox et al. 2007, *BMC Medical Genetics*; [4] Hinney et al. 2007, *PLoS One*; [5] Liu et al. 2008, *Human Molecular Genetics*; [6] Loos et al. 2008, *Nature Genetics*; [7] Thorleifsson et al. 2009, *Nature Genetics*; [8] Willer et al. 2009, *Nature Genetics*; [9] Meyere et al. 2009 *Nature Genetics*; [10] Cotsapas et al. 2009, *Human Molecular Genetics*; [11] Lindgren et al. 2009 *PLoS Genetics*; [12] Heard-Costa et al. 2009, *PLoS Genetics*; [13] Johansson et al. 2009, *Obesity*; [14] Liu et al. 2010, *Twin Research and Human Genetics*; [15] Shcerag et al. 2010, *PLoS Genetics*; Speliotes et al. 2010, *Nature Genetics*. LD Blocks were defined using an R^2 threshold of 0.95. Genes are reported within 100 kb of any seed SNP. Italicized genes fall outside the 100kb range, but contain SNPs in LD with a block seed. GWAS are indicated as replicating a block if they reported a SNP in LD at R^2≥0.95 with a block seed or proxy as associated with an obesity-related phenotype at $p<1x10^{-5}$ in either their discovery or combined discovery and replication samples.

Block 2.2: (**seeds**) rs10173167, rs10188334, rs10189761, rs10190052, rs10193244, rs11127484, rs11127485, rs11127491, rs12714414, rs12714415, rs12992154, rs12995480, rs13007080,

rs13007086, rs13012571, rs13021737, rs1320331, rs1320336, rs1320337, rs1320338, rs13386517, rs13386627, rs13386964, rs13388043, rs13393304, rs13396935, rs13397165, rs13401686, rs13415094, rs2860323, rs2867108, rs2867109, rs2867110, rs2867112, rs2867113, rs2867122, rs2867125, rs2903492, rs2947411, rs4423631, rs4452188, rs4613321, rs4854344, rs4854348, rs4854349, rs5017300, rs5017303, rs6711012, rs6719518, rs6719980, rs6725549, rs6728726, rs6731348, rs6731688, rs6732471, rs6734363, rs6743060, rs6744646, rs6744653, rs6752470, rs6755502, rs7561317, rs7567570, rs7570198, rs7571957, rs7574359, rs7576624, rs7576635, rs7585056, rs7604609, rs7608050, rs939582, rs939583

Block 2.3: (**seeds**) rs2867123, (**proxies**) rs10173167, rs10188334, rs10189761, rs10190052, rs10193244, rs11127484, rs11127485, rs11127491, rs12714414, rs12714415, rs12992154, rs12995480, rs13007080, rs13007086, rs13012571, rs13021737, rs1320331, rs1320336, rs1320337, rs1320338, rs13386517, rs13386627, rs13386964, rs13388043, rs13393304, rs13396935, rs13397165, rs13401686, rs13415094, rs2860323, rs2867108, rs2867109, rs2867110, rs2867112, rs2867113, rs2867122, rs2867123, rs2867125, rs2903492, rs4423631, rs4452188, rs4613321, rs4854344, rs4854348, rs4854349, rs5017300, rs5017303, rs6711012, rs6719518, rs6719980, rs6725549, rs6728726, rs6731348, rs6731688, rs6732471, rs6734363, rs6743060, rs6744646, rs6744653, rs6752470, rs6755502, rs7561317, rs7567570, rs7570198, rs7571957, rs7574359, rs7576624, rs7576635, rs7585056, rs7604609, rs7608050, rs939582, rs939583

Supplementary Table 2.5. Effect Sizes for Genetic Risk Scores Created Using the 3-Stage Approach and the Best-Guess and Top-Hits Approaches. To measure BMI effect sizes for the GRSs, we estimated Pearson correlations (r) from separate linear regressions of BMI on each of the GRSs. To measure obesity effect sizes for the GRSs, we estimated odds ratios (OR) from separate logistic regressions of obesity on each of the GRSs. Regression models were adjusted for age (linear and quadratic terms), gender, the age-gender interaction, and the ARIC Study Centers where data were collected. In Panel A, the Best-Guess GRS was based on the GRS published by Li and colleagues [62] and the Top-Hits GRS was based on the GRS published by Peterson and colleagues [61]. In Panel B, the Best Guess GRS was based on the full set of obesity- and BMI-associated SNPs listed in the NHGRI GWAS Catalog and the Top-Hits GRS was based on the GRS published by Speliotes and colleagues [89]. ***p<0.001. Comparison of effect sizes using the seemingly unrelated regression method [112] indicated that effect sizes for the 3 GRSs in Panel A were not statistically different from one another (p-value for difference >0.10 for all), but that among the GRSs in Panel B, the 3-stage approach performed better than the Best-Guess and Top-Hits GRSs (p<0.05 for all). However, our sample had only 40% power to detect effect size differences of r=0.01 / OR=1.01, so this result should be interpreted with caution.

		Effect Sizes	
		BMI	Obesity
Approach to GRS Construction	SNPs	Pearson Correlation (r)	Odds Ratio [95% CI]
Panel A. GRSs Constructed from Results of 9 GWAS Published by December 31, 2008			
3-Stage	28	0.08***	1.08 [1.06-1.10]
Best-Guess	12	0.08***	1.08 [1.06-1.11]
Top-Hits	59	0.06***	1.07 [1.04-1.09]
Panel B. GRSs Constructed from Results of the Full Set of 16 GWAS			
3-Stage	57	0.11***	1.12 [1.10-1.15]
Best-Guess	97	0.10***	1.11 [1.09-1.13]
Top-Hits	32	0.10***	1.10 [1.08-1.12]

Supplementary Table 2.6. Analysis of Loci with Multiple Tag SNPs. * "Lead SNP" is underlined; "Worst-associated SNP" is italicized; Test statistics and effect sizes were estimated in linear regression models of BMI adjusted for demographics and geography. "Lead SNPs" and "Worst-associated SNPs" were determined from the test statistics for the individual SNPs. Effect sizes were compared using the seemingly unrelated regressions method [112].

Locus	ARIC SNPs Tagging LD Blocks in Genic Region	Minimum R^2 Among Tag SNPs	Effect Size (Pearson's r) p-value for comparison with lead SNP		
			Lead SNP	Worst-Associated SNP	Mean Number of BMI-Increasing Alleles
Chr 2 *TMEM18*	*rs10189761* , rs2867123, rs4854345	0.94	0.027	0.023 p=0.276	0.025 p=0.371
Chr 3 ETV5/DGKG	rs12516728, *rs9863591*	0.85	0.007	<0.001 p=0.721	0.018 p=0.427
Chr 11 BDNF	rs10501087, *rs7103411* , rs6265, rs11030108	0.86	0.027	0.022 p=0.124	0.026 p=0.485
Chr 11 MTCH	rs12419692, *rs3817334*	0.77	0.020	0.019 p=0.871	0.020 p=0.878
Chr 16 FTO	*rs1477196* , rs17817288, rs1121980, rs9922047, rs9939973, rs9940128, rs9941349, rs7193144, rs7203521, rs9939609, rs8050136, rs9930506	0.40	0.072	0.034 p<0.001	0.068 p=0.104
Chr 18 MC4R	*rs476828* , rs1673518, rs17782313, rs11663816, rs11665563, rs12969709, rs12970134	0.25	0.026	0.019 p=0.158	0.025 p=0.062
Chr 19 KCDT15	*rs29942* , rs11084753	0.58	0.010	0.009 p=0.879	0.009 p=0.913

Supplementary Table 2.7. SNPs Included in the Obesity Genetic Risk Score.

								White Participants, n=8,210-8,8,286				Black Participants, n=2,402-2,442			
Chr	Nearby Gene	Tag SNP	GWAS Replications	BMI-Increasing Allele in GWAS	Test Alelle	Other Allele	Effect-Size Weight	Test Allele Frequency	Per Allele Change in BMI	p-value	Direction of Association Inconsistent with GWAS	Test Allele Frequency	Per Allele Change in BMI	p-value	Direction of Association Inconsistent with GWAS
1	NEGR1	rs2815752	3	Major	G	A	0.13	38%	-0.259	0.001		45%	-0.071	0.673	
	TNNI3K	rs1514175	1	Minor	A	G	0.07	43%	-0.001	0.985		68%	-0.091	0.608	X
	PTBP2	rs1555543	2	Major	A	C	0.06	42%	-0.128	0.086		57%	-0.031	0.855	
	SEC16B	rs543874	2	Minor	G	A	0.22	20%	0.341	0.000		25%	0.335	0.095	
2	FANCL	rs759250	1	Minor	A	G	0.10	29%	0.036	0.656		8%	-0.242	0.475	X
	LRP1B	rs2121279	1	Minor	T	C	0.08	14%	0.234	0.032		3%	-0.253	0.651	X
	TMEM18	rs2867123	5	Major	G	C	0.30	17%	-0.237	0.018		12%	0.022	0.935	X
	RBJ	rs10182181	1	Minor	G	A	0.14	46%	0.117	0.117		84%	0.758	0.001	
3	CADM2	rs12714640	1	Minor	A	C	0.10	19%	0.278	0.003		6%	0.006	0.987	
	ETV5/DGKG	rs1516728	2	Major	T	A	0.11	23%	-0.060	0.489		52%	-0.098	0.565	
4	GNPDA2	rs12641981	2	Minor	T	C	0.18	43%	0.088	0.238		23%	0.103	0.602	
	SLC39A8	rs13114738	1	Minor	T	C	0.13	8%	0.506	4.15E-04		1%	-1.583	0.008	X
5	POC5 FLJ35779	rs10057967	1	Major	C	T	0.10	37%	-0.227	0.003		49%	0.128	0.435	X
	ZNF608	rs6864049	1	Minor	G	A	0.07	46%	-0.189	0.012	X	19%	-0.463	0.033	X
6	TFAP2B	rs734597	3	Minor	A	G	0.13	17%	0.382	1.21E-04		9%	0.030	0.920	
9	LING02 LRRN6C	rs1412235	1	Minor	C	G	0.11	31%	0.003	0.970		16%	0.365	0.111	
	LMX1B	rs867559	2	Minor	G	A	0.24	20%	0.088	0.339		32%	0.025	0.889	
11	RPL27A	rs2028882	1	Major	C	A	0.06	50%	-0.065	0.375		66%	0.116	0.515	X
	BDNF	rs10501087	2	Major	C	T	0.18	21%	-0.223	0.013		7%	-0.521	0.181	
	MTCH2	rs12419692	2	Minor	A	C	0.05	36%	0.146	0.059		9%	0.012	0.968	
12	BDCDIN3D, FAIM2	rs7138803	3	Minor	A	G	0.12	38%	0.164	0.033		17%	0.100	0.650	
13	MTIF3, GRF3A	rs1475219	1	Minor	C	T	0.09	21%	0.262	0.004		22%	-0.099	0.632	X
14	PRKD1	rs1440983	1	Minor	A	G	0.15	5%	0.266	0.129		23%	0.156	0.449	
	NRXN3	rs7144011	2	Minor	T	G	0.13	22%	0.165	0.064		24%	0.164	0.428	
15	MAP2K5	rs28670272	1	Major	G	A	0.13	23%	-0.212	0.014		41%	0.005	0.977	X
16	GPR5B	rs11639988	1	Major	G	A	0.17	15%	0.006	0.952	X	24%	-0.262	0.194	
	ATXN2L, TUFM, SH2B1	rs12443881	3	Minor	T	C	0.15	39%	-0.005	0.948	X	9%	-0.607	0.030	X
	FTO	rs9939609	11	Minor	A	T	0.38	41%	0.496	8.19E-11		48%	0.129	0.443	
18	MC4R	rs12970134	6	Minor	A	G	0.21	26%	0.209	0.012		13%	0.057	0.822	
19	KCTD15	rs11084753	3	Major	A	G	0.04	33%	-0.071	0.371		36%	0.197	0.270	X
	QPCTL	rs11083779	1	Major	C	T	0.07	4%	-0.227	0.196		11%	-0.267	0.294	
	ZC3H4 TMEM160	rs7250850	1	Major	G	C	0.09	29%	-0.174	0.032		80%	-0.343	0.124	

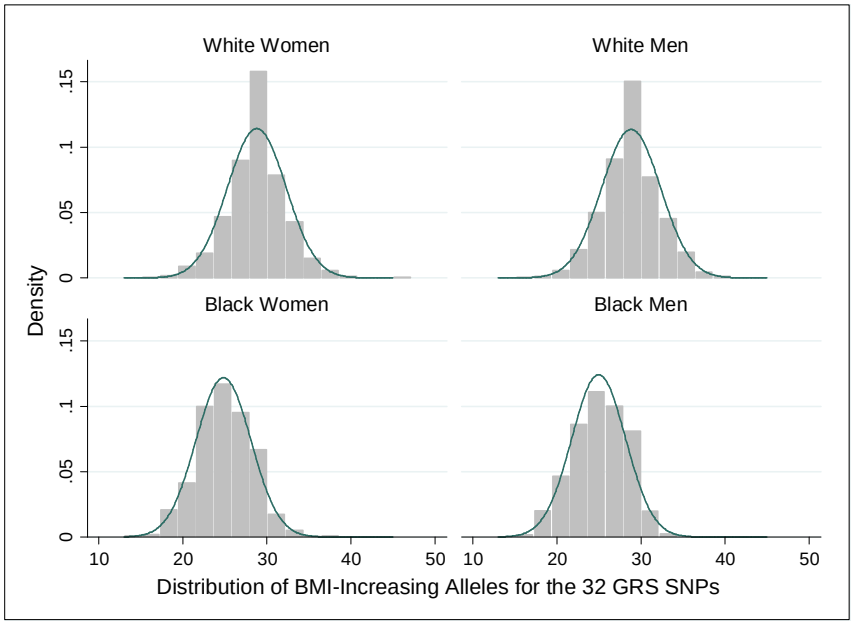
Supplementary Table 2.7 Footnote: GWAS replications include GWAS reporting any SNP in any LD block tagged by the SNP as obesity-associated at $p<1x10^{-5}$ in the discovery or combined discovery and replication samples. Test allele and other allele are reported from the positive strand. Effect-size weights were obtained from [89] for all SNPs with the exception of rs867559, for which the effect size weight was obtained from [90]. Allele frequencies and per-allele effects are reported based on all participants in the analysis sample. Per-allele effects were estimated from linear regressions of BMI on SNP genotype (number of minor alleles), adjusted for demographics and geography. P-values are reported based on heteroskedasticity robust standard errors.

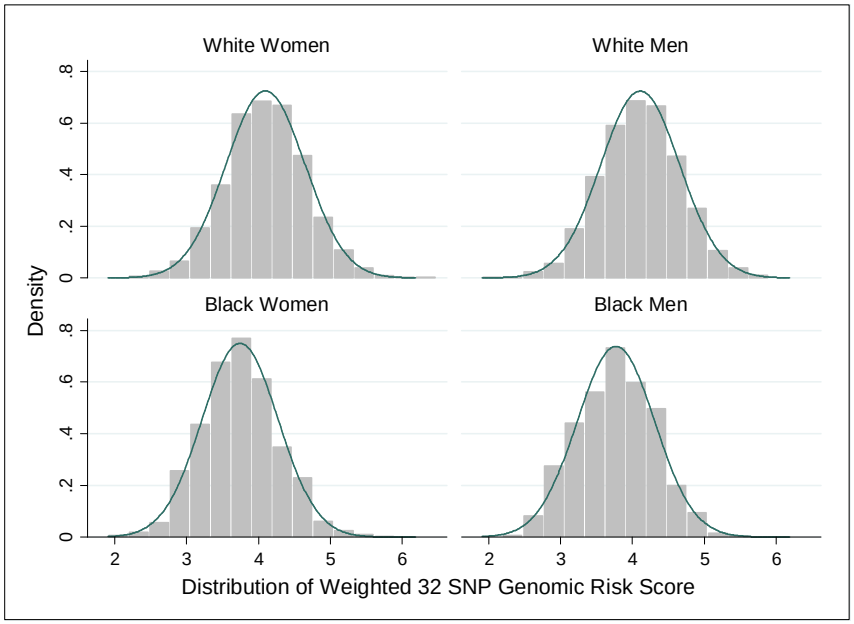
Supplementary Table 2.8. Educational Attainment of White and African American ARIC Participants. Educational attainment was ascertained via self-report at the first ARIC visit. Distributions of BMI-increasing alleles for the 32 obesity GRS SNPs were comparable across educational strata in African Americans and whites (p>0.10 for all comparisons).

Highest Level of Schooling	Percent of Visit 1 Sample	
None/ Grade School	5%	19%
Some High School	11%	21%
High School Graduate	36%	22%
Vocational School	9%	7%
College	30%	18%
Graduate/ Professional School	9%	14%

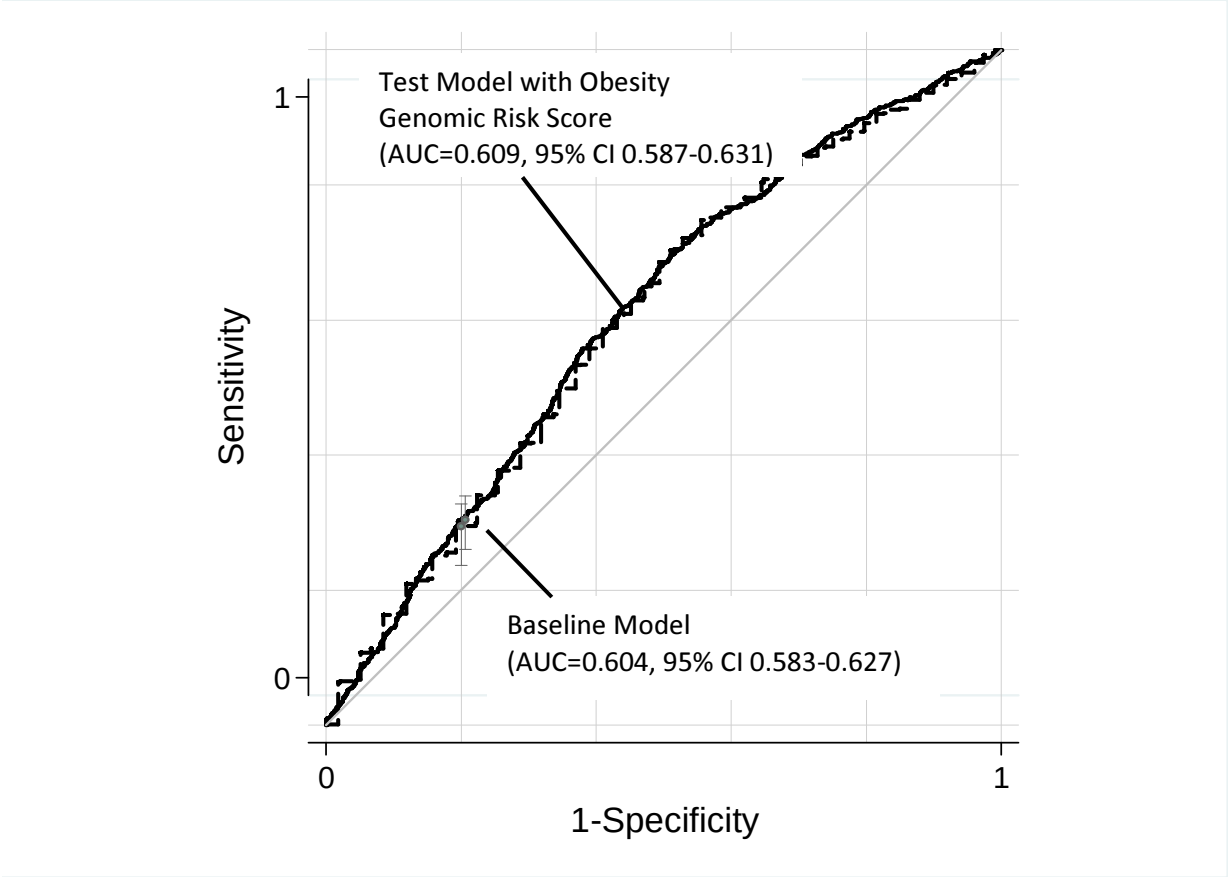
Supplementary Table 2.9. Predictiveness of Model-Based Risk Scores With and Without The Obesity Genetic Risk Score. (m1-5) denote separate models used to estimate risk scores for BMI and obesity. Risk scores were predicted values from linear regression of BMI and predicted probabilities from probit regressions of obesity. The first model, m1, includes measures of age, sex, and ARIC Study Center where data were collected. The regression model was specified to include linear and quadratic terms for age and a product term modeling interaction between age and sex. The simple genetic risk assessment (SNPs in *FTO* and downstream of *MC4R*) is a component of the weighted obesity genomic risk score. Thus, model m3 contains all of the information in model m2 as well as information from the remaining 30 SNPs included in the GRS. The 5 categories of socioeconomic status were modeled as dichotomous variables and were allowed to vary by sex in their relationship with obesity and BMI. Values of R^2 were estimated using linear regression models adjusted for demographic and geographic information. Percentile-based confidence intervals were generated using the bootstrap method. AUCs and percentile-based confidence intervals were estimated from ROC curves constructed for predicted values generated using a probit regression model and were adjusted for the ARIC Study Center where data were collected using Pepe's method [93, 95]. IDIs and test statistics were estimated only for comparisons of models m3 and m2 and models m5 and m4 using Pencina's Method [130]. IDIs for comparisons of models m2 and m3 with model m1 are identical to those reported for the respective obesity risk measures in Table 4 of the article.

		White ARIC Participants (n=8,286)			Black ARIC Participants (n=2,442)		
Model	Model Components	R^2 (95% CI)	AUC (95% CI)	IDI (p-value)	R^2 (95% CI)	AUC (95% CI)	IDI (p-value)
(m1)	Demographic & Geographic Information	3.20%	0.526		5.17%	0.604	
(m2)	m1 + Simple Genetic Risk Assessment	3.88%	0.550		5.35%	0.607	
(m3)	m1 + Weighted GRS	4.88%	0.574		5.52%	0.609	
	Change in predictiveness with addition of the weighted GRS	1.00% (0.006-0.014)	0.024 (0.012-0.036)	0.006 (7.81E-13)	0.17% (-0.001-0.005)	0.002 (-0.005-0.009)	0.001 (0.055)
(m4)	m1 + Socioeconomic Status	4.70%	0.550		7.70%	0.643	
(m5)	m4 + Weighted GRS	6.20%	0.586		7.92%	0.645	
	Change in predictiveness with addition of the weighted GRS	1.50% (0.010-0.020)	0.036 (0.023-0.050)	0.010 (5.46E-19)	0.22% (-0.001-0.006)	0.002 (-0.003-0.008)	0.002 (0.012)





Supplementary Figure 2.1. Distributions of BMI Increasing Alleles for the 32 GRS SNPs and the Weighted Obesity Genomic Risk Score Among White and African American ARIC Participants. Variance of the obesity genomic risk scores (GRS) was similar among women and men within ethnicity ($p>0.15$ for both samples), but was greater among whites as compared to African Americans ($p<0.001$) according to Brown and Forsythe's [131] test for equality of variances.



Supplementary Figure 2.2. Receiver Operating Characteristic Curves for Obesity Among African American ARIC Participants (n=2,442). Baseline Model = gender, age (quadratic), gender x age interaction, ARIC study center; Test Model = baseline model + weighted obesity genomic risk score. ROC Curves were constructed using predicted values from probit regressions of obesity (BMI≥30) on the model terms. Delta AUC (AUC_{Test}-$AUC_{Baseline}$) = 0.005, 95% CI -0.005-0.015, p=0.30. Delta Partial AUC at 80% specificity=0, 95% CI -0.004-0.004, p=0.97. AUCs, partial AUCs, and delta AUCs were estimated using Pepe's method [93, 95].

APPENDIX B

Supplementary Materials to Chapter 3

Supplementary Methods

Construction of the Obesity Genetic Risk Score. We selected single nucleotide polymorphisms (SNPs) for investigation that were associated with an obesity-related phenotype at a threshold of $p<1x10^{-5}$ in GWAS of European-descent individuals. We grouped the selected SNPs into "linkage disequilibrium (LD) blocks" using a linkage threshold of $R^2 \geq 0.95$ and data from the International HapMap Consortium's CEU sample.[155] We retained LD blocks that included a SNP associated with an obesity-related phenotype at $p<1x10^{-8}$ in ≥1 GWAS or that included SNPs associated with an obesity-related phenotype at $p<1x10^{-5}$ in ≥2 GWAS. This analysis yielded 32 LD blocks. We selected one tag SNP from each LD block to include in the GRS. To construct the GRS, we weighted the obesity-associated alleles for each GRS SNP by the effect size reported for the SNP or its closest LD proxy in meta-analyses of BMI GWAS.[89, 90] We then summed the weighted counts of obesity-associated alleles for each SNP to compute the GRS. An additive model was assumed on the basis of prior research documenting additive contributions to BMI for many of the GRS SNPs.[61, 62]

Supplemental Table 3.1. Single Nucleotide Polymorphisms Included in the Genetic Risk Score. Alleles are reported from the forward strand. *Nearest gene is reported for the locus identified in the meta-analysis of body mass index (BMI) GWAS by the GIANT Consortium.[89] GWAS effect sizes are the per-allele change in BMI estimated in meta-analyses of BMI GWAS by the GIANT Consortium and Thorleifsson and colleagues.[90] The following 3 SNPs failed quality controls in the Dunedin sample and were not included in the genetic risk score: rs11083779 near *QPCTL*; rs12641981 near *GNPDA2*; rs2121279 near *LRP1B*.

Chr	Nearest Gene*	rs	Alleles	BMI-Increasing Allele	Frequency of BMI-Increasing Allele	GWAS Effect-Size for BMI
1	NEGR1	rs2568958	A/G	A	60%	0.13
	TNNI3K	rs1514177	C/G	G	43%	0.07
	PTBP2	rs11165643	C/T	T	63%	0.06
	SEC16B	rs10913469	C/T	C	21%	0.21
2	TMEM18	rs7567570	C/T	C	82%	0.31
	ADCY3, RBJ	rs10182181	A/G	G	51%	0.14
	FANCL	rs887912	A/G	A	29%	0.10
3	CADM2	rs7640855	A/G	A	20%	0.10
	ETV5	rs7647305	C/T	C	79%	0.12
4	SLC39A8	rs13107325	C/T	T	8%	0.19
5	FLJ35779	rs2112347	G/T	T	65%	0.10
	ZNF608	rs6864049	A/G	G	55%	0.07
6	TFAP2B	rs2206277	A/G	A	18%	0.13
9	LRRN6C	rs1412235	C/G	C	31%	0.11
	LMX1B	rs867559	A/G	G	21%	0.24
11	STK33, RPL27A	rs4929949	C/T	C	52%	0.06
	BDNF	rs6265	A/G	G	52%	0.18
	MTCH2	rs10838738	A/G	G	34%	0.05
12	BCDIN3, FAIM2	rs7138803	A/G	A	36%	0.12
13	MTIF3	rs1475219	C/T	C	20%	0.09
14	PRKD1	rs11847697	C/T	T	3%	0.17
	NRXN3	rs10150332	C/T	C	23%	0.13
15	MAP2K5	rs2241423	A/G	G	78%	0.13
16	GPRC5B	rs12446554	G/T	G	87%	0.17
	SH2B1	rs4788102	A/G	A	39%	0.15
	FTO	rs9939609	A/T	A	36%	0.38
18	MC4R	rs921971	C/T	C	28%	0.21
19	KCTD15	rs29941	C/T	C	67%	0.06
	ZC3H4, TMEM160	rs3810291	A/G	A	68%	0.09

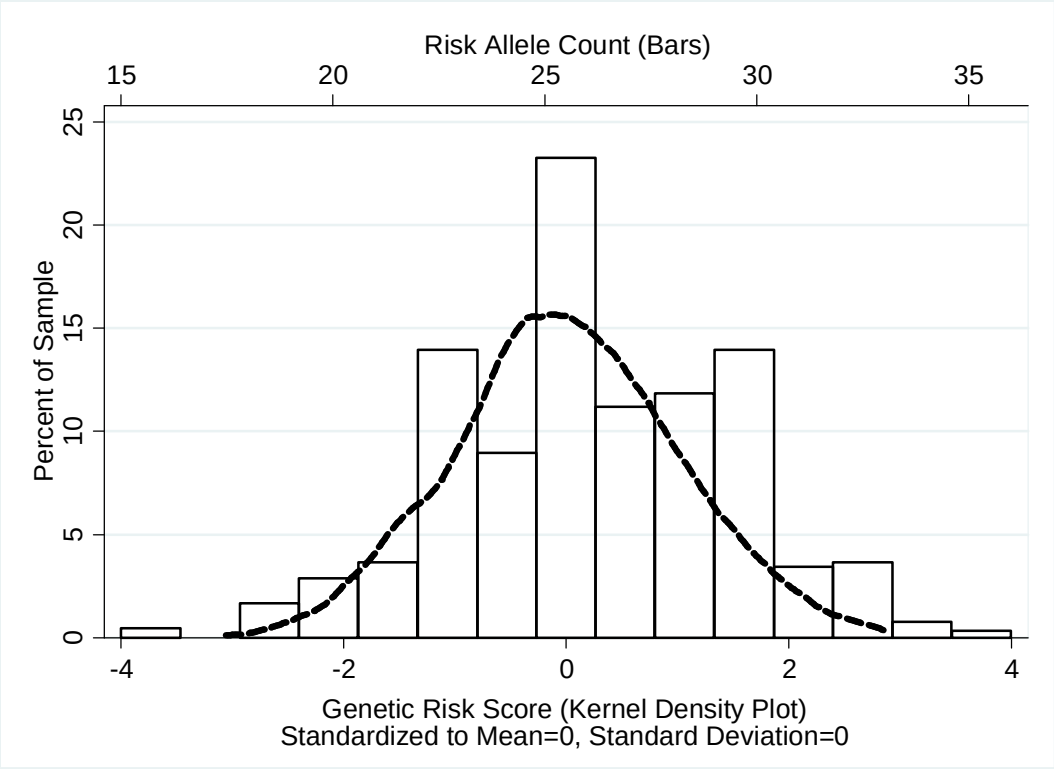
Supplemental Table 3.2. The Genetic Risk Score and the Family History Score have Independent Effects on Growth and Obesity Risk. Panel A presents bivariate effect sizes for the genetic risk score and the family history score from the life course growth model and obesity prediction models. Panel B presents the independent effects of the genetic risk score and the family history score on life course growth and obesity risk. Independent effects were estimated from multivariate growth models (life course growth) and multivariate Poisson regression models (obesity). The genetic risk score and the family history score were standardized to have means of 0 and standard deviations of 1 for analyses. All analyses were adjusted for sex.

	Life Course Growth			Obesity			
	Model Intercept	Childhood Slope	Adulthood Slope	Teens	20s	30s	Chronic
Panel A. Bivariate Associations							
		Beta/ p-value		Relative Risk (95% Confidence Interval)			
Genome Risk Score	0.38	0.03	0.02	1.42	1.37	1.23	1.37
	p<0.001	p<0.001	p=0.014	(1.10, 1.83)	(1.13, 1.67)	(1.08, 1.39)	(1.09, 1.73)
Family History Score	0.63	0.05	0.04	1.63	1.72	1.49	1.83
	p<0.001	p<0.001	p<0.001	(1.36, 1.95)	(1.51, 1.97)	(1.35, 1.63)	(1.58, 2.13)
Panel B. Independent Associations							
		Beta/ p-value		Relative Risk (95% Confidence Interval)			
Genome Risk Score	0.31	0.02	0.01	1.31	1.27	1.17	1.26
	p<0.001	p=0.004	p=0.065	(1.01, 1.70)	(1.04, 1.55)	(1.03, 1.32)	(1.00, 1.59)
Family History Score	0.60	0.05	0.04	1.58	1.67	1.46	1.78
	p<0.001	p<0.001	p<0.001	(1.31, 1.90)	(1.46, 1.92)	(1.32, 1.61)	(1.52, 2.09)

Supplemental Table 3.3. Indirect effects of the genetic risk score on adult obesity outcomes mediated through birth-3 weight gain and the adiposity rebound. Indirect effects were estimated using the structural equation described by MacKinnon & Dwyer[140] implemented with Poisson regression models. Indirect effect estimates were exponentiated to compute risk ratios. Indirect effect estimates for the adiposity rebound reflect the combined indirect effects of age and BMI at adiposity rebound. Analyses were adjusted for sex. Confidence intervals were estimated from 5000 bootstrap repetitions.

	Obesity Outcome							
	Teens		20s		30s		Chronic	
Developmental Phenotype	Indirect Effect Expressed as a Relative Risk (95% Confidence Intervals)							
Birth-3 Weight Gain	1.06	(1.01, 1.12)	1.03	(1.00, 1.06)	1.01	(0.99, 1.04)	1.04	(1.01, 1.08)
Adiposity Rebound	1.32	(1.17, 1.52)	1.19	(1.11, 1.31)	1.12	(1.06, 1.19)	1.20	(1.11, 1.32)

Supplemental Figure 3.1. Distribution of the Genetic Risk Score. The transparent bars show the distribution of the count of risk alleles across the 29 SNPs included in the genetic risk score (i.e. before weights were applied). The kernel density plot shows the distribution of the weighted genetic risk score**.**



APPENDIX C

Supplementary Materials to Chapter 4

Supplemental Table 4.1. Single nucleotide polymorphisms (SNPs) included in the genetic risk score. Effect allele frequencies for the GWAS SNPs are based on the HapMap CEU sample (release 22 for SNPs rs12595538, rs8032771, and rs4105144; version 3 release 2 for SNPs rs16969968 and rs6495308). Linkage disequilibrium (LD) was obtained from 1000 Genomes project data for all SNPs except rs4105144. LD between this SNP and rs8102683 was obtained using HapMap Release 22 data. All allele frequency and linkage queries were run through the Broad Institute's SNAP tool (http://www.broadinstitute.org/mpg/snap/ldsearch.php). Effect allele frequencies for the SNPs genotyped in the Dunedin sample are based on n=880 European-descent study members.

Chr	Genes	GWAS SNP	Alleles	Effect Allele	Freq. (HapMap)	Dunedin SNP	LD with GWAS SNP	Alleles	Effect Allele	Freq. (Dunedin)
15	CHRNA5, CHRNA3, CHRNB4	rs16969968	A/G	A	39%	rs10519203	0.93	A/G	G	34%
		rs6495308	C/T	T	80%	rs4887069	1.00	A/G	A	79%
	ADAMTS7, MORF4L1	rs12595538	A/T	A	62%	rs7164529	0.90	A/G	G	61%
		rs8032771	A/G	A	52%	rs11072810	0.97	C/T	T	50%
19	EGLN2	rs7937	C/T	T	55%	rs7937	1.00	C/T	T	57%
	CYP2A6	rs4105144	A/G	G	74%	rs8102683	0.87	C/T	C	73%

Supplemental Table 4.2. Associations between genetic risk and clinical phenotypes of smoking behavior are mediated by developmental phenotypes of rapid progression from smoking initiation to heavy smoking. Indirect, direct, and total effects were estimated from the structural equation described by MacKinnon and Dwyer implemented using the methods described by Preacher and colleagues.[140, 142, 143] Percentile-based 95% confidence intervals were estimated from 1,000 bootstrap repetitions. Developmental phenotypes were early conversion to daily smoking (by age 15 years) and rapid progression to heavy smoking (by age 18 years). Both developmental phenotypes were associated with the latent adult smoking problems factor and with the individual clinical phenotypes ($p<0.001$ for all). Collectively, early conversion to daily smoking and rapid progression to heavy smoking explained 23% of the variance in the latent smoking problems factor.

	Total Effect of Genetic Risk	Direct (un-mediated) Effect of Genetic Risk	Indirect Effect of Genetic Risk Mediated Through Developmental Phenotypes	Proportion of Total Effect Accounted for by the Indirect Effect
	B /[95% CI] / p-value	B /[95% CI] / p-value	B /[95% CI] / p-value	%
Latent Adult Smoking Problems Factor				
	0.15 [0.05-0.26] p=4.33E-03	0.03 [-0.04-0.10] p=4.02E-01	0.12 [0.05-0.20] p=1.79E-03	81%

REFERENCES

1. Lander ES. Initial impact of the sequencing of the human genome. *Nature.* (2011) 470:187-197.

2. Khoury MJ, Gwinn M, Bowen MS, Dotson WD. Beyond Base Pairs to Bedside: A Population Perspective on How Genomics Can Improve Health. *Am J Public Health.* (2011).

3. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A.* (2009) 106:9362-9367.

4. Pearson TA, Manolio TA. How to interpret a genome-wide association study. *JAMA.* (2008) 299:1335-1344.

5. Matise TC, Ambite JL, Buyske S, et al. The next PAGE in understanding complex traits: Design for the analysis of Population Architecture Using Genetics and Epidemiology (PAGE) Study. *Am J Epidemiol.* (2011).

6. Khoury MJ, Feero WG, Reyes M, et al. The genomic applications in practice and prevention network. *Genet Med.* (2009) 11:488-494.

7. Ioannidis JPA. Personalized genetic prediction: Too limited, too expensive, or too soon? *Ann Intern Med.* (2009) 150:139-141.

8. Goddard KA, Robitaille J, Dowling NF, et al. Health-related direct-to-consumer genetic tests: a public health assessment and analysis of practices related to Internet-based tests for risk of thrombosis. *Public Health Genomics.* (2009) 12:92-104.

9. Janssens A, Gwinn M, Bradley LA, Oostra BA, van Duijn CM, Khoury MJ. A critical appraisal of the scientific basis of commercial genomic profiles used to assess health risks and personalize health interventions. *Am J Hum Genet.* (2008) 82:593-599.

10. Dietz WH. Overweight in childhood and adolescence. *N Engl J Med.* (2004) 350:855-857.

11. DiFranza JR, Richmond JB. Let the children be heard: lessons from studies of the early onset of tobacco addiction. *Pediatrics.* (2008) 121:623-624.

12. Plomin R, DeFries JC, McClearn GE, McGuffin P. *Behavioral Genetics*. 5 ed. New York: Worth Publishers; 2008.

13. Yang WJ, Kelly T, He J. Genetic epidemiology of obesity. *Epidemiol Rev.* (2007) 29:49-61.

14. Goldman D, Oroszi G, Ducci F. The genetics of addictions: uncovering the genes. *Nat Rev Genet.* (2005) 6:521-532.

15. Maher B. Personal genomes: The case of the missing heritability. *Nature.* (2008) 456:18-21.

16. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature.* (2009) 461:747-753.

17. Sullivan P. Don't give up on GWAS. *Mol Psychiatry.* (2012) 17:2-3.

18. Khoury MJ, McCabe LL, McCabe ERB. Genomic medicine - Population screening in the age of genomic medicine. *N Engl J Med.* (2003) 348:50-58.

19. Olstad DL, McCargar L. Prevention of overweight and obesity in children under the age of 6 years. *Applied Physiology Nutrition and Metabolism-Physiologie Appliquee Nutrition Et Metabolisme.* (2009) 34:551-570.

20. De Bellis MD. Developmental traumatology: a contributory mechanism for alcohol and substance use disorders. *Psychoneuroendocrinology.* (2002) 27:155-170.

21. Melvin CL, Dolan-Mullen P, Windsor RA, Whiteside HP, Goldenberg RL. Recommended cessation counselling for pregnant women who smoke: a review of the evidence. *Tob Control.* (2000) 9:80-84.

22. Hesketh KD, Campbell KJ. Interventions to Prevent Obesity in 0-5 Year Olds: An Updated Systematic Review of the Literature. *Obesity.* (2010) 18:S27-S35.

23. Flynn MAT, McNeil DA, Maloff B, et al. Reducing obesity and related chronic disease risk in children and youth: a synthesis of evidence with 'best practice' recommendations. *Obes Rev.* (2006) 7:7-66.

24. Grimshaw GM, Stanton A. Tobacco cessation interventions for young people. *Cochrane Database Syst Rev.* (2006):CD003289.

25. Park ER, Japuntich SJ, Rigotti NA, et al. A snapshot of smokers after lung and colorectal cancer diagnosis. *Cancer.* (2012).

26. Marteau TM, Lerman CY. Genetic risk and behavioural change. *BMJ.* (2001) 322:1056-1059.

27. Grant RW, Hivert M, Pandiscio JC, Florez JC, Nathan DM, Meigs JB. The clinical application of genetic testing in type 2 diabetes: a patient and physician survey. *Diabetologia.* (2009) 52:2299-2305.

28. Collins F. Has the revolution arrived? *Nature.* (2010) 464:674-675.

29. Gniady J. Regulating direct-to-consumer genetic testing: Protecting the consumer without quashing a medical revolution. *Fordham Law Rev.* (2008) 76:2429-2475.

30. Hunter DJ, Khoury MJ, Drazen JM. Letting the genome out of the bottle - Will we get our wish? *N Engl J Med.* (2008) 358:105-107.

31. Khoury MJ, Gwinn M, Burke W, Bowen S, Zimmern R. Will Genomics widen or help heal the schism between medicine and public health? *Am J Prev Med.* (2007) 33:310-317.

32. Khoury MJ, Gwinn M, Yoon PW, Dowling N, Moore CA, Bradley L. The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention? *Genet Med.* (2007) 9:665-674.

33. Burke W, Laberge AM, Press N. Debating Clinical Utility. *Public Health Genomics.* (2010) 13:215-223.

34. Hunter DJ, Kraft P. Drinking from the fire hose--statistical issues in genomewide association studies. *N Engl J Med.* (2007) 357:436-439.

35. Gibson G. Rare and common variants: twenty arguments. *Nat Rev Genet.* (2011) 13:135-145.

36. Bennett SN, Caporaso N, Fitzpatrick AL, et al. Phenotype harmonization and cross-study collaboration in GWAS consortia: the GENEVA experience. *Genet Epidemiol.* (2011) 35:159-173.

37. Lazzeroni LC, Ray A. The cost of large numbers of hypothesis tests on power, effect size and sample size. *Mol Psychiatry.* (2012) 17:108-114.

38. Manolio TA. Cohort studies and the genetics of complex disease. *Nat Genet.* (2009) 41:5-6.

39. Manolio TA, Bailey-Wilson JE, Collins FS. Opinion - Genes, environment and the value of prospective cohort studies. *Nat Rev Genet.* (2006) 7:812-820.

40. Ioannidis JP, Trikalinos TA, Khoury MJ. Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. *Am J Epidemiol.* (2006) 164:609-614.

41. Senese LC, Almeida ND, Fath AK, Smith BT, Loucks EB. Associations Between Childhood Socioeconomic Position and Adulthood Obesity. *Epidemiol Rev.* (2009) 31:21-51.

42. Galea S, Nandi A, Vlahov D. The social epidemiology of substance use. *Epidemiol Rev.* (2004) 26:36-52.

43. Chopra M, Darnton-Hill I. Tobacco and obesity epidemics: not so different after all? *BMJ.* (2004) 328:1558-1560.

44. The ARIC investigators. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. *Am J Epidemiol.* (1989) 129:687-702.

45. McMillen IC, Robinson JS. Developmental origins of the metabolic syndrome: Prediction, plasticity, and programming. *Physiol Rev.* (2005) 85:571-633.

46. Frongillo EA, Lampl M. Early identification of children at risk of developing obesity. *Arch Pediatr Adolesc Med.* (2011) 165:1043-1044.

47. Jackson K, M, Sher KJ, Rose RJ, Kapiro J. *Trajectories of tobacco use from adolescence to adulthood: Are the most informative phenotypes tobacco specific?* Bethesda: National Cancer Institute; 2009.

48. Bierut LJ. Genetic vulnerability and susceptibility to substance dependence. *Neuron.* (2011) 69:618-627.

49. McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet.* (2008) 9:356-369.

50. Amos CI, Spitz MR, Cinciripini P. Chipping away at the genetics of smoking behavior. *Nat Genet.* (2010) 42:366-368.

51. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *Am J Hum Genet.* (2010) 86:6-22.

52. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. *Science.* (2008) 322:881-888.

53. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet.* (2004) 74:106-120.

54. Sullivan PF. Spurious genetic associations. *Biol Psychiatry.* (2007) 61:1121-1126.

55. Kathiresan S, Melander O, Anevski D, et al. Polymorphisms associated with cholesterol and risk of cardiovascular events. *N Engl J Med.* (2008) 358:1240-1249.

56. Meigs JB, Shrader P, Sullivan LM, et al. Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N Engl J Med.* (2008) 359:2208-2219.

57. Ripatti S, Tikkanen E, Orho-Melander M, et al. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *Lancet.* (2010) 376:1393-1400.

58. Yang J, Manolio TA, Pasquale LR, et al. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet.* (2011) 43:519-525.

59. Wassenaar CA, Dong Q, Wei Q, Amos CI, Spitz MR, Tyndale RF. Relationship between CYP2A6 and CHRNA5-CHRNA3-CHRNB4 variation and smoking behaviors and lung cancer risk. *J Natl Cancer Inst.* (2011) 103:1342-1346.

60. Rodriguez S, Cook DG, Gaunt TR, Nightingale CM, Whincup PH, Day IN. Combined analysis of CHRNA5, CHRNA3 and CYP2A6 in relation to adolescent smoking behaviour. *J Psychopharm.* (2011) 25:915-923.

61. Peterson RE, Maes HH, Holmans P, et al. Genetic risk sum score comprised of common polygenic variation is associated with body mass index. *Hum Genet.* (2011) 129:221-230.

62. Li S, Zhao JH, Luan J, et al. Cumulative effects and predictive value of common obesity-susceptibility variants identified by genome-wide association studies. *Am J Clin Nutr.* (2010) 91:184-190.

63. Horne BD, Anderson JL, Carlquist JF, et al. Generating genetic risk scores from intermediate phenotypes for use in association studies of clinically significant endpoints. *Ann Hum Genet.* (2005) 69:176-186.

64. Shiffman D, Rowland CM, Sninsky JJ, Devlin JJ. Polymorphisms associated with coronary heart disease: Better by the score. *Curr Op Mol Ther.* (2006) 8:493-499.

65. Plomin R, Haworth CMA, Davis OSP. Common disorders are quantitative traits. *Nat Rev Genet.* (2009) 10:872-878.

66. Janssens A. Is the time right for translation research in genomics? *Eur J Epidemiol.* (2008) 23:707-710.

67. Khoury MJ, McBride CM, Schully SD, et al. The Scientific Foundation for Personal Genomics: Recommendations from a National Institutes of Health-Centers for Disease Control and Prevention Multidisciplinary Workshop. *Genet Med.* (2009) 11:559-567.

68. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* (2007) 17:1520-1528.

69. Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ. A navigator for human genome epidemiology. *Nat Genet.* (2008) 40:124-125.

70. Gibson G, Goldstein DB. Human genetics: The hidden text of genome-wide associations. *Curr Biol.* (2007) 17:R929-R932.

71. Orozco G, Barrett JC, Zeggini E. Synthetic associations in the context of genome-wide association scan signals. *Hum Mol Genet.* (2010) 19:R137-144.

72. Fisher R. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh.* (1918) 52:399-433.

73. Ogden CL, Carroll MD, Curtin LR, McDowell MA, Tabak CJ, Flegal KM. Prevalence of overweight and obesity in the United States, 1999-2004. *JAMA.* (2006) 295:1549-1555.

74. Adams KF, Schatzkin A, Harris TB, et al. Overweight, obesity, and mortality in a large prospective cohort of persons 50 to 71 years old. *N Engl J Med.* (2006) 355:763-778.

75. Allender S, Rayner M. The burden of overweight and obesity-related ill health in the UK. *Obes Rev.* (2007) 8:467-473.

76. Trogdon JG, Finkelstein EA, Hylands T, Dellea PS, Kamal-Bahl SJ. Indirect costs of obesity: a review of the current literature. *Obes Rev.* (2008) 9:489-500.

77. O'Rahilly S. Human genetics illuminates the paths to metabolic disease. *Nature.* (2009) 462:307-314.

78. McCarthy MI. Genomics, type 2 diabetes, and obesity. *N Engl J Med.* (2010) 363:2339-2350.

79. Hindorff LA, Junkins HA, Mehta JP, Manolio TA. A Catalog of Published Genome-Wide Association Studies. http://www.genome.gov/gwastudies/. Accessed April 30, 2010.

80. Folsom AR, Chambless LE, Ballantyne CM, et al. An assessment of incremental coronary risk prediction using C-reactive protein and other novel risk markers: the atherosclerosis risk in communities study. *Arch Intern Med.* (2006) 166:1368-1373.

81. GENEVA ARIC Project. Quality Control Report for the ARIC GWAS database. dbGaP Accession # phs000090.v1.p1 ed. Bethesda, MD; 2009.

82. Psaty BM, O'Donnell CJ, Gudnason V, et al. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circ Cardiovasc Genet.* (2009) 2:73-80.

83. Korn JM, Kuruvilla FG, McCarroll SA, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet.* (2008) 40:1253-1260.

84. Lyssenko V, Jonsson A, Almgren P, et al. Clinical Risk Factors, DNA Variants, and the Development of Type 2 Diabetes. *N Engl J Med.* (2008) 359:2220-2232.

85. Morrison AC, Bare LA, Chambless LE, et al. Prediction of coronary heart disease risk using a genetic risk score: The atherosclerosis risk in communities study. *Am J Epidemiol.* (2007) 166:28-35.

86. Talmud PJ, Hingorani AD, Cooper JA, et al. Utility of genetic and non-genetic risk factors in prediction of type 2 diabetes: Whitehall II prospective cohort study. *BMJ.* (2010) 340:b4838.

87. Demirkan A, Penninx BW, Hek K, et al. Genetic risk profiles for depression and anxiety in adult and elderly cohorts. *Mol Psychiatry.* (2010) 16:773-783.

88. He MA, Cornelis MC, Franks PW, Zhang CL, Hu FB, Qi L. Obesity Genotype Score and Cardiovascular Risk in Women With Type 2 Diabetes Mellitus. *Arterioscler Thromb Vasc Biol.* (2010) 30:327-U370.

89. Speliotes EK, Willer CJ, Berndt SI, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet.* (2010) 42:937-948.

90. Thorleifsson G, Walters GB, Gudbjartsson DF, et al. Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nat Genet.* (2009) 41:18-24.

91. McGeechan K, Macaskill P, Irwig L, Liew G, Wong TY. Assessing new biomarkers and predictive models for use in clinical practice: a clinician's guide. *Arch Intern Med.* (2008) 168:2304-2310.

92. Mihaescu R, van Zitteren M, van Hoek M, et al. Improvement of risk prediction by genomic profiling: reclassification measures versus the area under the receiver operating characteristic curve. *Am J Epidemiol.* (2010) 172:353-361.

93. Janes H, Pepe MS. Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. *Biometrika.* (2009) 96:371-382.

94. Pepe MS, Cai TX, Longton G. Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics.* (2006) 62:221-229.

95. Pepe M, Longton G, Janes H. Estimation and Comparison of Receiver Operating Characteristic Curves. *Stata J.* (2009) 9:1.

96. Pencina MJ, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Stat Med.* (2008) 27:157-172.

97. Brown MB, Forsythe AB. Robust tests for equality of variances. *Journal of the American Statistical Association.* (1974) 69:364-367.

98. Hardy R, Wills AK, Wong A, et al. Life course variations in the associations between FTO and MC4R gene variants and body size. *Hum Mol Genet.* (2010) 19:545-552.

99. Ford ES, Mokdad AH. Epidemiology of obesity in the Western Hemisphere. *J Clin Endocrinol Metab.* (2008) 93:S1-8.

100. Drewnowski A. Obesity, diets, and social inequalities. *Nutr Rev.* (2009) 67 Suppl 1:S36-39.

101. Benjamin AM, Suchindran S, Pearce K, et al. Gene by sex interaction for measures of obesity in the framingham heart study. *Journal of Obesity.* (2011) 2011:329038.

102. Heid IM, Jackson AU, Randall JC, et al. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat Genet.* (2010) 42:949-960.

103. Evans JP, Meslin EM, Marteau TM, Caulfield T. Genomics. Deflating the genomic bubble. *Science.* (2011) 331:861-862.

104. Jorde LB, Wooding SP. Genetic variation, classification and 'race'. *Nature Genetics.* (2004) 36:S28-33.

105. Norris JM, Langefeld CD, Talbert ME, et al. Genome-wide association study and follow-up analysis of adiposity traits in Hispanic Americans: the IRAS Family Study. *Obesity.* (2009) 17:1932-1941.

106. McCarthy JJ, Meyer J, Moliterno DJ, Newby LK, Rogers WJ, Topol EJ. Evidence for substantial effect modification by gender in a large-scale genetic association study of the metabolic syndrome among coronary heart disease patients. *Hum Genet.* (2003) 114:87-98.

107. McCarthy JJ. Gene by sex interaction in the etiology of coronary heart disease and the preceding metabolic syndrome. *Nutrition, metabolism, and cardiovascular diseases : NMCD.* (2007) 17:153-161.

108. Chiu YF, Chuang LM, Kao HY, et al. Sex-specific genetic architecture of human fatness in Chinese: the SAPPHIRe Study. *Hum Genet.* (2010) 128:501-513.

109. Sovio U, Mook-Kanamori DO, Warrington NM, et al. Association between common variation at the FTO locus and changes in body mass index from infancy to late childhood: The complex nature of genetic association through growth and development. *PLoS Genet.* (2011) 7:e1001307.

110. Elks CE, Loos RJ, Sharp SJ, et al. Genetic markers of adult obesity risk are associated with greater early infancy weight gain and growth. *PLoS Med.* (2010) 7:e1000284.

111. Verzilli CJ, Stallard N, Whittaker JC. Bayesian modelling of multivariate quantitative traits using seemingly unrelated regressions. *Genet Epidemiol.* (2005) 28:313-325.

112. Baltagi BH. On Seemingly Unrelated Regressions with Error-Components. *Econometrica.* (1980) 48:1547-1551.

113. Royston P, Ambler G. *Multivariable fractional polynomials: Update.* College Station, Tx: Stata; 1999.

114. Royston P, Sauerbrei W. Stability of multivariable fractional polynomial models with selection of variables and transformations: a bootstrap investigation. *Stat Med.* (2003) 22:639-659.

115. Herbert A, Gerry NP, McQueen MB, et al. A common genetic variant is associated with adult and childhood obesity. *Science.* (2006) 312:279-283.

116. Frayling TM, Timpson NJ, Weedon MN, et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science.* (2007) 316:889-894.

117. Scuteri A, Sanna S, Chen WM, et al. Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet.* (2007) 3:1200-1210.

118. Fox CS, Heard-Costa N, Cupples LA, Dupuis J, Vasan RS, Atwood LD. Genome-wide association to body mass index and waist circumference: the Framingham Heart Study 100K project. *BMC Med Genet.* (2007) 8 Suppl 1:S18.

119. Hinney A, Nguyen TT, Scherag A, et al. Genome wide association (GWA) study for early onset extreme obesity supports the role of fat mass and obesity associated gene (FTO) variants. *Plos One.* (2007) 2:e1361.

120. Liu YJ, Liu XG, Wang L, et al. Genome-wide association scans identified CTNNBL1 as a novel gene for obesity. *Hum Mol Genet.* (2008) 17:1803-1813.

121. Loos RJ, Lindgren CM, Li S, et al. Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat Genet.* (2008) 40:768-775.

122. Willer CJ, Speliotes EK, Loos RJ, et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet.* (2009) 41:25-34.

123. Meyre D, Delplanque J, Chevre JC, et al. Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nat Genet.* (2009) 41:157-159.

124. Cotsapas C, Speliotes EK, Hatoum IJ, et al. Common body mass index-associated variants confer risk of extreme obesity. *Hum Mol Genet.* (2009) 18:3502-3507.

125. Lindgren CM, Heid IM, Randall JC, et al. Genome-Wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. *PLoS Genet.* (2009) 5:e1000508.

126. Heard-Costa NL, Zillikens MC, Monda KL, et al. NRXN3 is a novel locus for waist circumference: a genome-wide association study from the CHARGE Consortium. *PLoS Genet.* (2009) 5:e1000539.

127. Johansson A, Marroni F, Hayward C, et al. Linkage and genome-wide association analysis of obesity-related phenotypes: Association of weight with the MGAT1 gene. *Obesity.* (2010) 18:803-808.

128. Liu JZ, Medland SE, Wright MJ, et al. Genome-wide association study of height and body mass index in Australian twin families. *Twin Res Hum Genet.* (2010) 13:179-193.

129. Scherag A, Dina C, Hinney A, et al. Two new Loci for body-weight regulation identified in a joint analysis of genome-wide association studies for early-onset extreme obesity in French and german study groups. *PLoS Genet.* (2010) 6:e1000916.

130. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med.* (2008) 27:157-172; discussion 207-112.

131. Brown MB, Forsythe AB. Robust Tests for Equality of Variances. *J Am Stat Assoc.* (1974) 69:364-367.

132. Oken E, Gillman MW. Fetal origins of obesity. *Obes Res.* (2003) 11:496-506.

133. North KE, Graff M, Adair LS, et al. Genetic epidemiology of BMI and body mass change from adolescence to young adulthood. *Obesity.* (2010) 18:1474-1476.

134. Rolland-Cachera MF, Deheeger M, Maillot M, Bellisle F. Early adiposity rebound: causes and consequences for obesity in children and adults. *Int J Obes.* (2006) 30 Suppl 4:S11-17.

135. Taylor RW, Grant AM, Goulding A, Williams SM. Early adiposity rebound: review of papers linking this to subsequent obesity in children and adults. *Curr Op Clin Nutr Metab Care.* (2005) 8:607-612.

136. Janssen I, Katzmarzyk PT, Srinivasan SR, et al. Utility of childhood BMI in the prediction of adulthood disease: comparison of national and international references. *Obes Res.* (2005) 13:1106-1115.

137. Abdullah A, Wolfe R, Stoelwinder JU, et al. The number of years lived with obesity and the risk of all-cause and cause-specific mortality. *Int J Epidemiol.* (2011) 40:985-996.

138. Lobstein T, Baur L, Uauy R. Obesity in children and young people: a crisis in public health. *Obes Rev.* (2004) 5 Suppl 1:4-104.

139. Singer JD, Willett JB. *Applied Longitudinal Data Analysis*. New York: Oxford University Press; 2003.

140. Mackinnon DP, Dwyer JH. Estimating mediated effects in prevention studies. *Eval Rev.* (1993) 17:144-158.

141. Preacher KJ, Hayes AF. SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behav Res Methods Instrum Comput.* (2004) 36:717-731.

142. Preacher KJ, Hayes AF. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav Res Meth.* (2008) 40:879-891.

143. Preacher KJ, Kelley K. Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychol Methods.* (2011) 16:93-115.

144. *SAS* [computer program]. Version 9.2. Durham: SAS Institute Inc.; 2002-2008.

145. *Stata/MP for Windows* [computer program]. Version 11.1. College Station, TX: Stata Corp LP; 2010.

146. Eriksson J, Forsen T, Tuomilehto J, Osmond C, Barker D. Size at birth, childhood growth and obesity in adult life. *Int J Obes.* (2001) 25:735-740.

147. Li H, Stein AD, Barnhart HX, Ramakrishnan U, Martorell R. Associations between prenatal and postnatal growth and adult body size and composition. *Am J Clin Nutr.* (2003) 77:1498-1505.

148. Williams SM, Goulding A. Patterns of growth associated with the timing of adiposity rebound. *Obesity.* (2009) 17:335-341.

149. Whitaker KL, Jarvis MJ, Beeken RJ, Boniface D, Wardle J. Comparing maternal and paternal intergenerational transmission of obesity risk in a large population-based sample. *Am J Clin Nutr.* (2010) 91:1560-1567.

150. Eichler EE, Flint J, Gibson G, et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet.* (2010) 11:446-450.

151. Gibson G. Hints of hidden heritability in GWAS. *Nat Genet.* (2010) 42:558-560.

152. Qiu J, Hayden EC. Genomics sizes up. *Nature.* (2008) 451:234.

153. Andersson EA, Pilgaard K, Pisinger C, et al. Do gene variants influencing adult adiposity affect birth weight? A population-based study of 24 loci in 4,744 Danish individuals. *PLoS One.* (2010) 5:e14190.

154. Kilpelainen TO, den Hoed M, Ong KK, et al. Obesity-susceptibility loci have a limited influence on birth weight: a meta-analysis of up to 28,219 individuals. *Am J Clin Nutr.* (2011) 93:851-860.

155. International HapMap Project. International HapMap Project Genotype Data, Release 1. 3 ed; 2010.

156. Centers for Disease Control and Prevention (CDC). Smoking-attributable mortality, years of potential life lost, and productivity losses--United States, 2000-2004. *MMWR Morb Mortal Wkly Rep.* (2008) 57:1226-1228.

157. Current cigarette smoking prevalence among working adults--United States, 2004-2010. *MMWR Morb Mortal Wkly Rep.* (2011) 60:1305-1309.

158. Joseph AM, Fu SS, Lindgren B, et al. Chronic disease management for tobacco dependence: a randomized, controlled trial. *Arch Intern Med.* (2011) 171:1894-1900.

159. Carpenter MJ, Hughes JR, Gray KM, Wahlquist AE, Saladin ME, Alberg AJ. Nicotine therapy sampling to induce quit attempts among smokers unmotivated to quit: a randomized clinical trial. *Arch Intern Med.* (2011) 171:1901-1907.

160. Carson KV, Brinn MP, Labiszewski NA, Esterman AJ, Chang AB, Smith BJ. Community interventions for preventing smoking in young people. *Cochrane Database of Systematic Reviews.* (2011).

161. Xian H, Scherrer JF, Madden PA, et al. The heritability of failed smoking cessation and nicotine withdrawal in twins who smoked and attempted to quit. *Nicotine Tobacco Res.* (2003) 5:245-254.

162. Thorgeirsson TE, Gudbjartsson DF, Surakka I, et al. Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. *Nat Genet.* (2010) 42:448-453.

163. Liu JZ, Tozzi F, Waterworth DM, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet.* (2010) 42:436-440.

164. Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet.* (2010) 42:441-447.

165. Sullivan PF, Kendler KS. The genetic epidemiology of smoking. *Nicotine Tobacco Res.* (1999) 1 Suppl 2:S51-57; discussion S69-70.

166. Baker TB, Weiss RB, Bolt D, et al. Human neuronal acetylcholine receptor A5-A3-B4 haplotypes are associated with multiple nicotine dependence phenotypes. *Nicotine Tobacco Res.* (2009) 11:785-796.

167. Munafo MR, Johnstone EC, Walther D, Uhl GR, Murphy MF, Aveyard P. CHRNA3 rs1051730 genotype and short-term smoking cessation. *Nicotine Tobacco Res.* (2011) 13:982-988.

168. Lerman C, Jepson C, Wileyto EP, et al. Genetic variation in nicotine metabolism predicts the efficacy of extended-duration transdermal nicotine therapy. *Clin Pharmacol Ther.* (2010) 87:553-557.

169. King DP, Paciga S, Pickering E, et al. Smoking cessation pharmacogenetics: analysis of varenicline and bupropion in placebo-controlled clinical trials. *Neuropsychopharmacology.* (2012) 37:641-650.

170. Baker TB, Conti DV, Moffitt TE, Caspi A. *The nicotine dependence phenotype: Translating theoretical perspectives and extant data into recommendations for genetic mapping.* Bethesda: National Cancer Insititute; 2009.

171. Gelernter J. Developmental perspective on the role of genes in smoking risk. *Biol Psychiatry.* (2011) 69:616-617.

172. Bierut LJ. Nicotine dependence and genetic variation in the nicotinic receptors. *Drug Alcohol Depend.* (2009) 104 Suppl 1:S64-69.

173. Breslau N, Peterson EL. Smoking cessation in young adults: age at initiation of cigarette smoking and other suspected influences. *Am J Public Health.* (1996) 86:214-220.

174. Buchmann AF, Blomeyer D, Jennen-Steinmetz C, et al. Early smoking onset may promise initial pleasurable sensations and later addiction. *Addict Biol.* (2011).

175. Kandel DB, Hu MC, Griesler PC, Schaffran C. On the development of nicotine dependence in adolescence. *Drug Alcohol Depend.* (2007) 91:26-39.

176. Dierker L, Swendsen J, Rose J, He J, Merikangas K. Transitions to Regular Smoking and Nicotine Dependence in the Adolescent National Comorbidity Survey (NCS-A). *Ann Behav Med.* (2011).

177. Ducci F, Kaakinen M, Pouta A, et al. TTC12-ANKK1-DRD2 and CHRNA5-CHRNA3-CHRNB4 influence different pathways leading to smoking behavior from adolescence to mid-adulthood. *Biol Psychiatry.* (2011) 69:650-660.

178. Berrettini WH, Doyle GA. The CHRNA5-A3-B4 gene cluster in nicotine addiction. *Mol Psychiatry.* (2011).

179. Malaiyandi V, Sellers EM, Tyndale RF. Implications of CYP2A6 genetic variation for smoking behaviors and nicotine dependence. *Clin Pharmacol Ther.* (2005) 77:145-158.

180. Kubota T, Nakajima-Taniguchi C, Fukuda T, et al. CYP2A6 polymorphisms are associated with nicotine dependence and influence withdrawal symptoms in smoking cessation. *Pharmacogenomics J.* (2006) 6:115-119.

181. Weiss RB, Baker TB, Cannon DS, et al. A candidate gene approach identifies the CHRNA5-A3-B4 region as a risk factor for age-dependent nicotine addiction. *PLoS Genet.* (2008) 4:e1000125.

182. Audrain-McGovern J, Al Koudsi N, Rodriguez D, Wileyto EP, Shields PG, Tyndale RF. The role of CYP2A6 in the emergence of nicotine dependence in adolescents. *Pediatrics.* (2007) 119:e264-274.

183. O'Loughlin J, Paradis G, Kim W, et al. Genetically decreased CYP2A6 and the risk of tobacco dependence: a prospective study of novice smokers. *Tob Control.* (2004) 13:422-428.

184. Huang S, Cook DG, Hinks LJ, et al. CYP2A6, MAOA, DBH, DRD4, and 5HT2A genotypes, smoking behaviour and cotinine levels in 1518 UK adolescents. *Pharmacogenet Genomics.* (2005) 15:839-850.

185. Johnson EO, Chen LS, Breslau N, et al. Peer smoking and the nicotinic receptor genes: an examination of genetic and environmental risks for nicotine dependence. *Addiction.* (2010) 105:2014-2022.

186. Saccone NL, Culverhouse RC, Schwantes-An TH, et al. Multiple independent loci at chromosome 15q25.1 affect smoking quantity: a meta-analysis and comparison with lung cancer and COPD. *PLoS Genet.* (2010) 6.

187. Milne BJ, Moffitt TE, Crump R, et al. How should we construct psychiatric family history scores? A comparison of alternative approaches from the Dunedin Family Health History Study. *Psychol Med.* (2008) 38:1793-1802.

188. Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics.* (1988) 44:1049-1060.

189. Prentice RL, Williams BJ, Peterson AV. On the Regression-Analysis of Multivariate Failure Time Data. *Biometrika.* (1981) 68:373-379.

190. Hughes JR, Callas PW. Definition of a quit attempt: a replication test. *Nicotine Tobacco Res.* (2010) 12:1176-1179.

191. Broms U, Wedenoja J, Largeau MR, et al. Analysis of detailed phenotype profiles reveals CHRNA5-CHRNA3-CHRNB4 gene cluster association with several nicotine dependence traits. *Nicotine Tobacco Res.* (2012).

192. Benowitz NL. Nicotine addiction. *N Engl J Med.* (2010) 362:2295-2303.

193. Saccone NL, Schwantes-An TH, Wang JC, et al. Multiple cholinergic nicotinic receptor genes affect nicotine dependence risk in African and European Americans. *Genes, Brain, and Behavior.* (2010) 9:741-750.

194. Boardman JD, Blalock CL, Pampel FC, Hatemi PK, Heath AC, Eaves LJ. Population composition, public policy, and the genetics of smoking. *Demography.* (2011) 48:1517-1533.

195. Fiore MC, Baker TB. Stealing a march in the 21st century: accelerating progress in the 100-year war against tobacco addiction in the United States. *Am J Public Health.* (2009) 99:1170-1175.

196. Rose G. Sick individuals and sick populations. *Int J Epidemiol.* (1985) 14:32-38.

197. Smith GD. Epidemiology, epigenetics and the 'Gloomy Prospect': embracing randomness in population health research and practice. *Int J Epidemiol.* (2011) 40:537-562.

198. Guttmacher AE, Porteous ME, McInerney JD. Educating health-care professionals about genetics and genomics. *Nat Rev Genet.* (2007) 8:151-157.

199. Heatherton TF, Kozlowski LT, Frecker RC, Fagerstrom KO. The Fagerstrom Test for Nicotine Dependence: a revision of the Fagerstrom Tolerance Questionnaire. *Br J Addict.* (1991) 86:1119-1127.

200. Fagerstrom KO. Measuring degree of physical dependence to tobacco smoking with reference to individualization of treatment. *Addict Behav.* (1978) 3:235-241.

201. Bierut LJ, Madden PA, Breslau N, et al. Novel genes identified in a high-density genome wide association study for nicotine dependence. *Hum Mol Genet.* (2007) 16:24-35.

202. Saccone SF, Hinrichs AL, Saccone NL, et al. Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. *Hum Mol Genet.* (2007) 16:36-49.

203. Caspi A, Moffitt TE, Thornton A, et al. The life history calendar: A research and clinical assessment method for collecting retrospective event-history data. *Int J Method Psych.* (1996) 6:101-114.

204. Belli RF, Smith LM, Andreski PM, Agrawal S. Methodological comparisons between CATI event history calendar and standardized conventional questionnaire instruments. *Public Opin Q.* (2007) 71:603-622.

205. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Optimal designs for two-stage genome-wide association studies. *Genet Epidemiol.* (2007) 31:776-788.

206. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet.* (2010) 11:499-511.

207. Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. *Science.* (2002) 296:2225-2229.

208. Ntzani EE, Liberopoulos G, Manolio TA, Ioannidis JP. Consistency of genome-wide associations across major ancestral groups. *Hum Genet.* (2011).

209. Janssens A, Aulchenko YS, Elefante S, Borsboom G, Steyerberg EW, van Duijin CM. Predictive testing for complex diseases using multiple genes: Fact or fiction? *Genet Med.* (2006) 8:395-400.

210. Ioannidis JP, Panagiotou OA. Comparison of effect sizes associated with biomarkers reported in highly cited individual articles and in subsequent meta-analyses. *JAMA.* (2011) 305:2200-2210.

211. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. *PLoS Biol.* (2010) 8:e1000294.

212. Anderson CA, Soranzo N, Zeggini E, Barrett JC. Synthetic associations are unlikely to account for many common disease genome-wide association signals. *PLoS Biol.* (2011) 9:e1000580.

213. Wray NR, Purcell SM, Visscher PM. Synthetic associations created by rare variants do not explain most GWAS results. *PLoS Biol.* (2011) 9:e1000579.

214. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A.* (2012) 109:1193-1198.

215. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era--concepts and misconceptions. *Nat Rev Genet.* (2008) 9:255-266.

216. Yoon PW, Scheuner MT, Khoury MJ. Research priorities for evaluating family history in the prevention of common chronic diseases. *Am J Prev Med.* (2003) 24:128-135.

217. United States Department of Health and Human Services. Surgeon General's Family Health History Innitiative. http://www.hhs.gov/familyhistory/. Accessed March 1, 2010.

218. Yoon PW, Scheuner MT, Peterson-Oehlke KL, Gwinn M, Faucett A, Khoury MJ. Can family history be used as a tool for public health and preventive medicine? *Genet Med.* (2002) 4:304-310.

219. Evans DM, Visscher PM, Wray NR. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum Mol Genet.* (2009) 18:3525-3531.

220. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet.* (2011) 88:294-305.

221. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk of complex disease. *Current Opinion in Genetics & Development.* (2008) 18:257-263.

222. Yang J, Benyamin B, McEvoy BP, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet.* (2010) 42:565-569.

223. Purcell SM, Wray NR, Stone JL, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature.* (2009) 460:748-752.

224. Painter JN, Anderson CA, Nyholt DR, et al. Genome-wide association study identifies a locus at 7p15.2 associated with endometriosis. *Nat Genet.* (2011) 43:51-54.

225. Braveman PA, Kumanyika S, Fielding J, et al. Health disparities and health equity: the issue is justice. *Am J Public Health.* (2011) 101 Suppl 1:S149-155.

226. Linder SH, Sexton K. Conceptual models for cumulative risk assessment. *Am J Public Health.* (2011) 101 Suppl 1:S74-81.

227. Mackenbach JP. What would happen to health inequalities if smoking were eliminated? *BMJ.* (2011) 342:d3460.

228. Sturm R. The effects of obesity, smoking, and drinking on medical problems and costs. *Health Aff (Millwood).* (2002) 21:245-253.

229. Wang Y, Beydoun MA. The obesity epidemic in the United States - Gender, age, socioeconomic, Racial/Ethnic, and geographic characteristics: A systematic review and meta-regression analysis. *Epidemiol Rev.* (2007) 29:6-28.

230. Barbeau EM, Krieger N, Soobader MJ. Working class matters: socioeconomic disadvantage, race/ethnicity, gender, and smoking in NHIS 2000. *Am J Public Health.* (2004) 94:269-278.

231. Link BG, Phelan J. Social conditions as fundamental causes of disease. *J Health Soc Behav.* (1995) Spec No:80-94.

232. Shanahan MJ, Hofer SM. Social context in gene-environment interactions: retrospect and prospect. *J Gerontol B Psychol Sci Soc Sci.* (2005) 60 Spec No 1:65-76.

233. Juster RP, McEwen BS, Lupien SJ. Allostatic load biomarkers of chronic stress and impact on health and cognition. *Neurosci Biobehav Rev.* (2010) 35:2-16.

234. Kelly S, Hertzman C, Daniels M. Searching for the biological pathways between stress and health. *Annu Rev Public Health.* (1997) 18:437-462.

235. Shonkoff JP, Boyce WT, McEwen BS. Neuroscience, molecular biology, and the childhood roots of health disparities: building a new framework for health promotion and disease prevention. *JAMA.* (2009) 301:2252-2259.

236. Caspi A, Moffitt TE. Gene-environment interactions in psychiatry: joining forces with neuroscience. *Nat Rev Neurosci.* (2006) 7:583-590.

237. Hariri AR. The neurobiology of individual differences in complex behavioral traits. *Annu Rev Neurosci.* (2009) 32:225-247.

238. Carnell S, Gibson C, Benson L, Ochner CN, Geliebter A. Neuroimaging and obesity: current knowledge and future directions. *Obes Rev.* (2012) 13:43-56.

239. Sinha R, Li CS. Imaging stress- and cue-induced drug and alcohol craving: association with relapse and clinical implications. *Drug Alcohol Rev.* (2007) 26:25-31.

240. Hartwell KJ, Johnson KA, Li X, et al. Neural correlates of craving and resisting craving for tobacco in nicotine dependent smokers. *Addict Biol.* (2011) 16:654-666.

241. Jasinska AJ, Chua HF, Ho SS, Polk TA, Rozek LS, Strecher VJ. Amygdala response to smoking-cessation messages mediates the effects of serotonin transporter gene variation on quitting. *Neuroimage.* (2012) 60:766-773.

242. Engelmann JM, Versace F, Robinson JD, et al. Neural substrates of smoking cue reactivity: A meta-analysis of fMRI studies. *Neuroimage.* (2011) 60:252-262.

243. Li S, Zhao JH, Luan J, et al. Physical activity attenuates the genetic predisposition to obesity in 20,000 men and women from EPIC-Norfolk prospective population study. *PLoS Med.* (2010) 7.